OXFORD STATISTICAL SCIENCE SERIES • 38

Time Series Analysis by State Space Methods

SECOND EDITION

J. Durbin and S. J. Koopman

OXFORD STATISTICAL SCIENCE SERIES

OXFORD STATISTICAL SCIENCE SERIES

For a full list of titles please visit
http://www.oup.co.uk/academic/science/maths/series/osss/

10. J.K. Lindsey: *Models for Repeated Measurements*
11. N.T. Longford: *Random Coefficient Models*
12. P.J. Brown: *Measurement, Regression, and Calibration*
13. Peter J. Diggle, Kung-Yee Liang, and Scott L. Zeger: *Analysis of Longitudinal Data*
14. J.I. Ansell and M.J. Phillips: *Practical Methods for Reliability Data Analysis*
15. J.K. Lindsey: *Modelling Frequency and Count Data*
16. J.L. Jensen: *Saddlepoint Approximations*
17. Steffen L. Lauritzen: *Graphical Models*
18. A.W. Bowman and A. Azzalini: *Applied Smoothing Techniques for Data Analysis*
19. J.K. Lindsey: *Models for Repeated Measurements, Second Edition*
20. Michael Evans and Tim Swartz: *Approximating Integrals via Monte Carlo and Deterministic Methods*
21. D.F. Andrews and J.E. Stafford: *Symbolic Computation for Statistical Inference*
22. T.A. Severini: *Likelihood Methods in Statistics*
23. W.J. Krzanowski: *Principles of Multivariate Analysis: A User's Perspective, Revised Edition*
24. J. Durbin and S.J. Koopman: *Time Series Analysis by State Space Methods*
25. Peter J. Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott L. Zeger: *Analysis of Longitudinal Data, Second Edition*
26. J.K. Lindsey: *Nonlinear Models in Medical Statistics*
27. Peter J. Green, Nils L. Hjort, and Sylvia Richardson: *Highly Structured Stochastic Systems*
28. Margaret Sullivan Pepe: *The Statistical Evaluation of Medical Tests for Classification and Prediction*
29. Christopher G. Small and Jinfang Wang: *Numerical Methods for Nonlinear Estimating Equations*
30. John C. Gower and Garmt B. Dijksterhuis: *Procrustes Problems*
31. Margaret Sullivan Pepe: *The Statistical Evaluation of Medical Tests for Classification and Prediction, Paperback*
32. Murray Aitkin, Brian Francis, and John Hinde: *Statistical Modelling in GLIM4, Second Edition*
33. Anthony C. Davison, Yadolah Dodge, N. Wermuth: *Celebrating Statistics: Papers in Honour of Sir David Cox on his 80th Birthday*
34. Anthony Atkinson, Alexander Donev, and Randall Tobias: *Optimum Experimental Designs, with SAS*
35. M. Aitkin, B. Francis, J. Hinde, and R. Darnell: *Statistical Modelling in R*
36. Ludwig Fahrmeir and Thomas Kneib: *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*
37. Raymond L. Chambers and Robert G. Clark: *An Introduction to Model-Based Survey Sampling with Applications*
38. J. Durbin and S.J. Koopman: *Time Series Analysis by State Space Methods, Second Edition*

Time Series Analysis by State Space Methods

Second Edition

J. Durbin
London School of Economics
and Political Science
and University College London

S. J. Koopman
Vrije Universiteit Amsterdam

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford. If furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and in certain other countries

First Edition published in 2001
Reprinted 2008
Second Edition published in 2012

Impression: 1

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Library of Congress Control Number: 2011945385

ISBN 978–0–19–964117–8

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

To Anne
JD

To my family
SJK

This page intentionally left blank

Preface to Second Edition

This edition is about 100 pages longer than the first edition. The main reason for this is our desire to make the treatment more comprehensive. For example, we provide a more extensive foundation of filtering and smoothing for which we present proofs for a Bayesian analysis and for linear unbiased estimation as well as for a classical analysis. Our treatments for the linear model are based on four lemmas from multivariate regression theory. We have completely rewritten the discussions on simulation smoothing methods and we have added sections on dynamic factor analysis and state smoothing algorithms, including the Whittle smoothing relations and the two filter formula for smoothing. For a selection of chapters we have added a final section with exercises.

Part II of this book is on the analysis of nonlinear and non-Gaussian state space models and has been completely rewritten. We have introduced treatments for approximate solutions to filtering and smoothing including the well-known extended Kalman filter but also a self-contained treatment on unscented filtering and smoothing which we have completed with further extensions. We also treat approximate methods based on data transformations and mode computations. The chapter on the importance sampling method is fully rewritten and it can now also treat cases where the importance density is not necessarily log concave. Another new chapter in the book is devoted to particle filtering. We derive the main results and algorithms including the bootstrap filter and the auxiliary particle filter. Different approaches for designing an effective particle filter are presented. Finally, we discuss Bayesian estimation for state space models in a new chapter which includes separate sections for posterior analysis for linear Gaussian models and for nonlinear and non-Gaussian models by importance sampling methods. A new section provides more details on Markov chain Monte Carlo methods.

The process of writing this second edition has taken us about four years. JD thanks Professor Andrew Chesher, the Director of the Center for Microdata Methods and Practice (CEMMAP) and Professor of Economics at University College London, for proposing him for an Honorary Professorship at University College, London and for granting him the position of Honorary Research Fellow which provided him with an office and secretarial support. SJK thanks István Barra, Falk Brauning, Jacques Commandeur, Drew Creal, Kai Ming Lee and Marius Ooms for reading parts of the manuscript and for providing the support to do the work. SJK is also indebted to his colleagues and collaborators John Aston, João Valle é Azevedo, Frits Bijleveld, Charles Bos, Angeles Carnero, Jurgen Doornik, Virginie Dordonnat, Marc Francke, Irma Hindrayanto, Lennart Hoogerheide, Paweł Janus, Borus Jungbacker, André Lucas, Michael Massmann, Albert Menkveld, Geert Mesters, Marcel Scharth, Bernd Schwaab, Aart de Vos, Sunčica Vujić, Michel van der Wel, Brian Wong and Eric Zivot for their general support. Finally, SJK thanks Andrew Chesher for his generous support and hospitality which enabled him to make a number of visits to University College London.

London, J.D.
Amsterdam, S.J.K.
August 2011

Preface to First Edition

This book presents a comprehensive treatment of the state space approach to time series analysis. The distinguishing feature of state space time series models is that observations are regarded as made up of distinct components such as trend, seasonal, regression elements and disturbance terms, each of which is modelled separately. The models for the components are put together to form a single model called a state space model which provides the basis for analysis. The techniques that emerge from this approach are very flexible and are capable of handling a much wider range of problems than the main analytical system currently in use for time series analysis, the Box-Jenkins ARIMA system.

The exposition is directed primarily towards students, teachers, methodologists and applied workers in time series analysis and related areas of applied statistics and econometrics. Nevertheless, we hope that the book will also be useful to workers in other fields such as engineering, medicine and biology where state space models are employed. We have made a special effort to make the presentation accessible to readers with no previous knowledge of state space methods. For the development of all important parts of the theory, the only mathematics required that is not elementary is matrix multiplication and inversion, while the only statistical theory required, apart from basic principles, is elementary multivariate normal regression theory.

The techniques that we develop are aimed at practical application to real problems in applied time series analysis. Nevertheless, a surprising degree of beauty is to be found in the elegant way that many of the results drop out; no doubt this is due to the Markovian nature of the models and the recursive structure of the calculations.

State space time series analysis began with the pathbreaking paper of Kalman (1960) and early developments of the subject took place in the field of engineering. The term ‘state space’ came from engineering, and although it does not strike a natural rapport in statistics and econometrics, we, like others, use it because it is strongly established. We assure beginners that the meaning of the word ‘state’ will become clear very quickly; however, the attachment of the word ‘space’ might perhaps remain mysterious to non-engineers.

The book is divided into two parts. Part I discusses techniques of analysis based on the linear Gaussian state space model; the methods we describe represent the core of traditional state space methodology together with some new developments. We have aimed at presenting a state-of-the-art treatment of time series methodology based on this state space model. Although the model has been studied extensively over the past forty years, there is much that is new in our treatment. We use the word ‘new’ here and below to refer to results derived from original research in our recently published papers or obtained for the first time in this book. In particular, this usage relates to the treatment of disturbance smoothing in Chapter 4, exact initial filtering and smoothing in Chapter 5, the univariate treatment of multivariate observations plus computing algorithms in Chapter 6, aspects of diffuse likelihood, the score vector, the EM algorithm and allowing for the effects of parameter estimation on estimates of variance in Chapter 7, and the use of importance sampling for Bayesian analysis in Chapter 8. The linear Gaussian model, often after transformation of the observations, provides an adequate basis for the analysis of many of the time series that are encountered in practice. However, situations occur where this model fails to provide an acceptable representation of the data. For example, in the study of road accidents when the numbers observed are small, the Poisson distribution gives an intrinsically better model for the behaviour of the data than the normal distribution. There is therefore a need to extend the scope of state space methodology to cover non-Gaussian observations; it is also important to allow for nonlinearities in the model and for heavy-tailed densities to deal with outliers in the observations and structural shifts in the state.

These extensions are considered in Part II of the book. The treatment given there is new in the sense that it is based partly on the methods developed in our Durbin and Koopman (2000) paper and partly on additional new material. The methodology is based on simulation since exact analytical results are not available for the problems under consideration. Earlier workers had employed Markov chain Monte Carlo simulation methods to study these problems; however, during the research that led to this book, we investigated whether traditional simulation methods based on importance sampling and antithetic variables could be made to work satisfactorily. The results were successful so we have adopted this approach throughout Part II. The simulation methods that we propose are computationally transparent, efficient and convenient for the type of time series applications considered in this book.

An unusual feature of the book is that we provide analyses from both classical and Bayesian perspectives. We start in all cases from the classical standpoint since this is the mode of inference within which each of us normally works. This also makes it easier to relate our treatment to earlier work in the field, most of which has been done from the classical point of view. We discovered as we developed simulation methods for Part II, however, that we could obtain Bayesian solutions by relatively straightforward extensions of the classical treatments. Since our views on statistical inference are eclectic and tolerant, and since we believe that methodology for both approaches should be available to applied workers, we are happy to include solutions from both perspectives in the book.

Both the writing of the book and doing the research on which it was based have been highly interactive and highly enjoyable joint efforts. Subject to this, there have naturally been differences in the contributions that the two authors have made to the work. Most of the new theory in Part I was initiated by SJK while most of the new theory in Part II was initiated by JD. The expository and literary styles of the book were the primary responsibility of JD while the illustrations and computations were the primary responsibility of SJK.

Our collaboration in this work began while we were working in the Statistics Department of the London School of Economics and Political Science (LSE). We were fortunate in having as departmental colleagues three distinguished workers in the field. Our main thanks are to Andrew Harvey who helped us in many ways and whose leadership in the development of the methodology of state space time series analysis at the LSE was an inspiration to us both. We also thank Neil Shephard for many fruitful discussions on various aspects of the statistical treatment of state space models and for his incisive comments on an earlier draft of this book. We are grateful to Piet de Jong for some searching discussions on theoretical points.

We thank Jurgen Doornik of Nuffield College, Oxford for his help over a number of years which has assisted in the development of the computer packages *STAMP* and SsfPack. SJK thanks Nuffield College, Oxford for its hospitality and the Royal Netherlands Academy of Arts and Sciences for its financial support while he was at CentER, Tilburg University.

The book was written in LaTeX using the MiKTeX system (http://www.miktex.org). We thank Jurgen Doornik for assistance in setting up this LaTeX system.

London J.D.
Amsterdam S.J.K.
November 2000

This page intentionally left blank

Contents

PART I THE LINEAR STATE SPACE MODEL

PART II NON-GAUSSIAN AND NONLINEAR STATE SPACE MODELS

This page intentionally left blank

1 Introduction

1.1 Basic ideas of state space analysis

State space modelling provides a unified methodology for treating a wide range of problems in time series analysis. In this approach it is assumed that the development over time of the system under study is determined by an unobserved series of vectors $\alpha_1, \dots, \alpha_n$, with which are associated a series of observations $y_1, \dots, y_n$; the relation between the α_t's and the y_t's is specified by the state space model. The main purpose of state space analysis is to infer the relevant properties of the α_t's from a knowledge of the observations $y_1, \dots, y_n$. Other purposes include forecasting, signal extraction and estimation of parameters. This book presents a systematic treatment of this approach to problems of time series analysis.

Our starting point when deciding the structure of the book was that we wanted to make the basic ideas of state space analysis easy to understand for readers with no previous knowledge of the approach. We felt that if we had begun the book by developing the theory step by step for a general state space model, the underlying ideas would be obscured by the complicated appearance of many of the formulae. We therefore decided instead to devote Chapter 2 of the book to a particularly simple example of a state space model, the local level model, and to develop as many as possible of the basic state space techniques for this model. Our hope is that this will enable readers new to the techniques to gain insights into the ideas behind state space methodology that will help them when working through the greater complexities of the treatment of the general case. With this purpose in mind, we introduce topics such as Kalman filtering, state smoothing, disturbance smoothing, simulation smoothing, missing observations, forecasting, initialisation, maximum likelihood estimation of parameters and diagnostic checking for the local level model. We present the results from both classical and Bayesian standpoints. We demonstrate how the basic theory that is needed for both cases can be developed from elementary results in regression theory.

1.2 Linear models

Before going on to develop the theory for the general model, we present a series of examples that show how the linear state space model relates to problems of practical interest. This is done in Chapter 3 where we begin by showing how structural time series models can be put into state space form. By structural time series models we mean models in which the observations are made up of trend, seasonal, cycle and regression components plus error. We go on to put Box–Jenkins ARIMA models into state space form, thus demonstrating that these models are special cases of state space models. Next we discuss the history of exponential smoothing and show how it relates to simple forms of state space and ARIMA models. We follow this by considering various aspects of regression with or without time-varying coefficients or autocorrelated errors. We also present a treatment of dynamic factor analysis. Further topics discussed are simultaneous modelling series from different sources, benchmarking, continuous time models and spline smoothing in discrete and continuous time. These considerations apply to minimum variance linear unbiased systems and to Bayesian treatments as well as to classical models.

Chapter 4 begins with a set of four lemmas from elementary multivariate regression which provides the essentials of the theory for the general linear state space model from both a classical and a Bayesian standpoint. These have the useful property that they produce the same results for Gaussian assumptions of the model and for linear minimum variance criteria where the Gaussian assumptions are dropped. The implication of these results is that we only need to prove formulae for classical models assuming normality and they remain valid for linear minimum variance and for Bayesian assumptions. The four lemmas lead to derivations of the Kalman filter and smoothing recursions for the estimation of the state vector and its conditional variance matrix given the data. We also derive recursions for estimating the observation and state disturbances. We derive the simulation smoother which is an important tool in the simulation methods we employ later in the book. We show that allowance for missing observations and forecasting are easily dealt with in the state space framework.

Computational algorithms in state space analyses are mainly based on recursions, that is, formulae in which we calculate the value at time $t+1$ from earlier values for $t, t-1, \ldots, 1$. The question of how these recursions are started up at the beginning of the series is called initialisation; it is dealt with in Chapter 5. We give a general treatment in which some elements of the initial state vector have known distributions while others are diffuse, that is, treated as random variables with infinite variance, or are treated as unknown constants to be estimated by maximum likelihood.

Chapter 6 discusses further computational aspects of filtering and smoothing and begins by considering the estimation of a regression component of the model and intervention components. It next considers the square root filter and smoother which may be used when the Kalman filter and smoother show signs of numerical instability. It goes on to discuss how multivariate time series can be treated as univariate series by bringing elements of the observational vectors into the system one at a time, with computational savings relative to the multivariate treatment in some cases. Further modifications are discussed where the observation vector is high-dimensional. The chapter concludes by discussing computer algorithms.

In Chapter 7, maximum likelihood estimation of parameters is considered both for the case where the distribution of the initial state vector is known and for the case where at least some elements of the vector are diffuse or are treated as fixed and unknown. The use of the score vector and the EM algorithm is discussed. The effect of parameter estimation on variance estimation is examined.

Up to this point the exposition has been based on the classical approach to inference in which formulae are worked out on the assumption that parameters are known, while in applications unknown parameter values are replaced by appropriate estimates. In Bayesian analysis the parameters are treated as random variables with a specified or a noninformative prior joint density which necessitates treatment by simulation techniques which are not introduced until Chapter 13. Chapter 13 partly considers a Bayesian analysis of the linear Gaussian model both for the case where the prior density is proper and for the case where it is noninformative. We give formulae from which the posterior mean can be calculated for functions of the state vector, either by numerical integration or by simulation. We restrict attention to functions which, for given values of the parameters, can be calculated by the Kalman filter and smoother.

In Chapter 8 we illustrate the use of the methodology by applying the techniques that have been developed to a number of analyses based on real data. These include a study of the effect of the seat belt law on road accidents in Great Britain, forecasting the number of users logged on to an Internet server, fitting acceleration against time for a simulated motorcycle accident and a dynamic factor analysis for the term structure of US interest rates.

1.3 Non-Gaussian and nonlinear models

Part II of the book extends the treatment to state space models which are not both linear and Gaussian. Chapter 9 illustrates the range of non-Gaussian and nonlinear models that can be analysed using the methods of Part II. This includes exponential family models such as the Poisson distribution for the conditional distribution of the observations given the state. It also includes heavy-tailed distributions for the observational and state disturbances, such as the t-distribution and mixtures of normal densities. Departures from linearity of the models are studied for cases where the basic state space structure is preserved. Financial models such as stochastic volatility models are investigated from the state space point of view.

Chapter 10 considers approximate methods for analysis of non-Gaussian and nonlinear models, that is, extended Kalman filter methods and unscented methods. It also discusses approximate methods based on first and second order Taylor expansions. We show how to calculate the conditional mode of the state given the observations for the non-Gaussian model by iterated use of the Kalman filter and smoother. We then find the linear Gaussian model with the same conditional mode given the observations.

The simulation techniques for exact handling of non-Gaussian and nonlinear models are based on importance sampling and are described in Chapter 11. We then find the linear Gaussian model with the same conditional mode given the observations. We use the conditional density of the state given the observations for an approximating linear Gaussian model as the importance density. We draw random samples from this density for the simulation using the simulation smoother described in Chapter 4. To improve efficiency we introduce two antithetic variables intended to balance the simulation sample for location and scale.

In Chapter 12 we emphasise the fact that simulation for time series can be done sequentially, that is, instead of selecting an entire new sample for each time point t, which is the method suggested in Section 12.2, we fix the sample at the values previously obtained at time $\ldots, t-2, t-1$, and choose a new value at time t only. New recursions are required for the resulting simulations. This method is called particle filtering.

In Chapter 13 we discuss the use of importance sampling for the estimation of parameters in Bayesian analysis for models of Part I and Part II. An alternative simulation technique is Markov chain Monte Carlo. We prefer to use importance sampling for the problems considered in this book but a brief description is given for comparative purposes.

We provide examples in Chapter 14 which illustrate the methods that have been developed in Part II for analysing observations using non-Gaussian and nonlinear state space models. The illustrations include the monthly number of van drivers killed in road accidents in Great Britain, outlying observations in quarterly gas consumption, the volatility of exchange rate returns and analysis of the results of the annual boat race between teams of the universities of Oxford and Cambridge.

1.4 Prior knowledge

Only basic knowledge of statistics and matrix algebra is needed in order to understand the theory in this book. In statistics, an elementary knowledge is required of the conditional distribution of a vector y given a vector x in a multivariate normal distribution; the central results needed from this area for much of the theory of the book are stated in the lemmas in Section 4.2. Little previous knowledge of time series analysis is required beyond an understanding of the concepts of a stationary time series and the autocorrelation function. In matrix algebra all that is needed are matrix multiplication and inversion of matrices, together with basic concepts such as rank and trace.

1.5 Notation

Although a large number of mathematical symbols are required for the exposition of the theory in this book, we decided to confine ourselves to the standard

English and Greek alphabets. The effect of this is that we occasionally need to use the same symbol more than once; we have aimed however at ensuring that the meaning of the symbol is always clear from the context. We present below a list of the main conventions we have employed.

- The same symbol 0 is used to denote zero, a vector of zeros or a matrix of zeros.
- The symbol I_k denotes an identity matrix of dimension k.
- We use the generic notation $p(\cdot)$, $p(\cdot, \cdot)$, $p(\cdot|\cdot)$ to denote a probability density, a joint probability density and a conditional probability density.
- If x is a random vector with μ and variance matrix V and which is not necessarily normal, we write $x \sim (\mu, V)$.
- If x is a random vector which is normally distributed with mean vector μ and variance matrix V, we write $x \sim \mathrm{N}(\mu, V)$.
- If x is a random variable with the chi-squared distribution with ν degrees of freedom, we write $x \sim \chi^2_\nu$.
- We use the same symbol $\mathrm{Var}(x)$ to denote the variance of a scalar random variable x and the variance matrix of a random vector x.
- We use the same symbol $\mathrm{Cov}(x, y)$ to denote the covariance between scalar random variables x and y, between a scalar random variable x and a random vector y, and between random vectors x and y.
- The symbol $\mathrm{E}(x|y)$ denotes the conditional expectation of x given y; similarly for $\mathrm{Var}(x|y)$ and $\mathrm{Cov}(x, y|z)$ for random vectors x, y and z.
- The symbol $\mathrm{diag}(a_1, \dots, a_k)$ denotes the $\ell \times \ell$ matrix with nonsingular matrix elements $a_1, \dots, a_k$ down the leading diagonal and zeros elsewhere where $\ell = \sum_{i=1}^{k} \mathrm{rank}(a_i)$.

1.6 Other books on state space methods

Without claiming complete coverage, we list here a number of books which contain treatments of state space methods.

First we mention three early books written from an engineering standpoint: Jazwinski (1970), Sage and Melsa (1971) and Anderson and Moore (1979). A later book from a related standpoint is Young (1984).

Books written from the standpoint of statistics and econometrics include Harvey (1989), who gives a comprehensive state space treatment of structural time series models together with related state space material, West and Harrison (1997), who give a Bayesian treatment with emphasis on forecasting, Kitagawa and Gersch (1996) and Kim and Nelson (1999). A complete Bayesian treatment for specific classes of time series models including the state space model is given by Frühwirth-Schnatter (2006). Fundamental and rigorous statistical treatments of classes of hidden Markov models, which include our nonlinear non-Gaussian state space model of Part II, is presented in Cappé, Moulines and Rydén

(2005). An introductory and elementary treatment of state space methods from a practioners' perspective is provided by Commandeur and Koopman (2007).

More general books on time series analysis and related topics which cover partial treatments of state space topics include Brockwell and Davis (1987) (39 pages on state space out of about 570), Chatfield (2003) (14 pages out of about 300), Harvey (1993) (48 pages out of about 300), Hamilton (1994) (37 pages on state space out of about 800 pages) and Shumway and Stoffer (2000) (112 pages out of about 545 pages). The monograph of Jones (1993) on longitudinal models has three chapters on state space (66 pages out of about 225). The book by Fahrmeir and Tutz (1994) on multivariate analysis based on generalised linear modelling has a chapter on state space models (48 pages out of about 420). Finally, the book by Teräsvirta, Tjostheim and Granger (2011) on nonlinear time series modelling has one chapter on the treatment of nonlinear state space models (32 pages out of about 500 pages).

Books on time series analysis and similar topics with minor treatments of state space analysis include Granger and Newbold (1986) and Mills (1993). We mention finally the book edited by Doucet, De Freitas and Gordon (2001) which contains a collection of articles on Monte Carlo (particle) filtering and the book edited by Akaike and Kitagawa (1999) which contains 6 chapters (88 pages) on illustrations of state space analysis out of a total of 22 chapters (385 pages).

1.7 Website for the book

We will maintain a website for the book at

`http://www.ssfpack.com/dkbook.html`

for data, code, corrections and other relevant information. We will be grateful to readers if they inform us about their comments and errors in the book so corrections can be placed on the site.

Part I

The linear state space model

In Part I we present a full treatment of the construction and analysis of linear state space models, and we discuss the software required for implementing the resulting methodology. We begin with a treatment of the local level model to serve as an introduction to the methodology. For the general case we show that the Gaussian systems, the minimum variance linear unbiased estimators and their Bayesian variants all lead to the same formulae. Initialisation, missing observations, forecasting and diagnostic checking are considered. Methods based on these models, possibly after transformation of the observations, are appropriate for a wide range of problems in practical time series analysis. We present illustrations of the applications of the methods to real series. Exercises are provided for a selection of chapters.

This page intentionally left blank

2 Local level model

2.1 Introduction

The purpose of this chapter is to introduce the basic techniques of state space analysis, such as filtering, smoothing, initialisation and forecasting, in terms of a simple example of a state space model, the local level model. This is intended to help beginners grasp the underlying ideas more quickly than they would if we were to begin the book with a systematic treatment of the general case. We shall present results from both the classical and Bayesian perspectives, assuming normality, and also from the standpoint of minimum variance linear unbiased estimation when the normality assumption is dropped.

A *time series* is a set of observations $y_1, \dots, y_n$ ordered in time. The basic model for representing a time series is the additive model

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \qquad t = 1, \dots, n. \tag{2.1}$$

Here, μ_t is a slowly varying component called the *trend*, γ_t is a periodic component of fixed period called the *seasonal* and ε_t is an irregular component called the *error* or *disturbance*. In general, the observation y_t and the other variables in (2.1) can be vectors but in this chapter we assume they are scalars. In many applications, particularly in economics, the components combine multiplicatively, giving

$$y_t = \mu_t \gamma_t \varepsilon_t. \tag{2.2}$$

By taking logs however and working with logged values model (2.2) reduces to model (2.1), so we can use model (2.1) for this case also.

To develop suitable models for μ_t and γ_t we need the concept of a *random walk*. This is a scalar series α_t determined by the relation $\alpha_{t+1} = \alpha_t + \eta_t$ where the η_t's are independent and identically distributed random variables with zero means and variances σ_η^2.

Consider a simple form of model (2.1) in which $\mu_t = \alpha_t$ where α_t is a random walk, no seasonal is present and all random variables are normally distributed. We assume that ε_t has constant variance σ_ε^2. This gives the model

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t, \qquad & \varepsilon_t &\sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \\ \alpha_{t+1} &= \alpha_t + \eta_t, \qquad & \eta_t &\sim \mathrm{N}\big(0, \sigma_\eta^2\big), \end{aligned} \tag{2.3}$$

for $t = 1, \dots, n$ where the ε_t's and η_t's are all mutually independent and are independent of α_1. This model is called the *local level model.* Although it has a simple form, this model is not an artificial special case and indeed it provides the basis for the analysis of important real problems in practical time series analysis; for example, the local level model provides the basis for our analysis of the Nile data that we start in Subsection 2.2.5. It exhibits the characteristic structure of state space models in which there is a series of unobserved values $\alpha_1, \dots, \alpha_n$, called the *states*, which represents the development over time of the system under study, together with a set of *observations* $y_1, \dots, y_n$ which are related to the α_t's by the state space model (2.3). The object of the methodology that we shall develop is to infer relevant properties of the α_t's from a knowledge of the observations $y_1, \dots, y_n$. The model (2.3) is suitable for both classical and Bayesian analysis. Where the ε_t's and the η_t's are not normally distributed we obtain equivalent results from the standpoint of minimum variance linear unbiased estimation.

We assume initially that $\alpha_1 \sim \mathrm{N}(a_1,P_1)$ where a_1 and P_1 are known and that σ^2_ε and σ^2_η are known. Since random walks are non-stationary the model is non-stationary. By non-stationary here we mean that distributions of random variables y_t and α_t depend on time t.

For applications of model (2.3) to real series, we need to compute quantities such as the mean of α_t given $y_1, \dots, y_{t-1}$ or the mean of α_t given $y_1, \dots, y_n$, together with their variances; we also need to fit the model to data by calculating maximum likelihood estimates of the parameters σ^2_ε and σ^2_η. In principle, this could be done by using standard results from multivariate normal theory as described in books such as Anderson (2003). In this approach the observations y_t generated by the local level model are represented as the $n \times 1$ vector Y_n such that

$$Y_n \sim \mathrm{N}(\mathbf{1}a_1, \Omega), \quad \text{with} \quad Y_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \Omega = \mathbf{1}\mathbf{1}'P_1 + \Sigma, \tag{2.4}$$

where the (i, j)th element of the $n \times n$ matrix Σ is given by

$$\Sigma_{ij} = \begin{cases} (i-1)\sigma^2_\eta, & i < j \\ \sigma^2_\varepsilon + (i-1)\sigma^2_\eta, & i = j, \qquad i, j = 1, \dots, n, \\ (j-1)\sigma^2_\eta, & i > j \end{cases} \tag{2.5}$$

which follows since the local level model implies that

$$y_t = \alpha_1 + \sum_{j=1}^{t-1} \eta_j + \varepsilon_t, \qquad t = 1, \dots, n. \tag{2.6}$$

Starting from this knowledge of the distribution of Y_n, estimation of conditional means, variances and covariances is in principle a routine matter using standard results in multivariate analysis based on the properties of the multivariate normal distribution. However, because of the serial correlation between the observations y_t, the routine computations rapidly become cumbersome as n increases. This naive approach to estimation can be improved upon considerably by using the filtering and smoothing techniques described in the next three sections. In effect, these techniques provide efficient computing algorithms for obtaining the same results as those derived by multivariate analysis theory. The remaining sections of this chapter deal with other important issues such as fitting the local level model and forecasting future observations.

2.2 Filtering

2.2.1 The Kalman filter

The object of *filtering* is to update our knowledge of the system each time a new observation y_t is brought in. We shall first develop the theory of filtering for the local level model (2.3) where the ε_t's and η_t's are assumed normal from the standpoint of classical analysis. Since in this case all distributions are normal, conditional joint distributions of one set of observations given another set are also normal. Let Y_{t-1} be the vector of observations $(y_1, \ldots, y_{t-1})'$ for $t = 2, 3, \ldots$ and assume that the conditional distribution of α_t given Y_{t-1} is $\mathrm{N}(a_t, P_t)$ where a_t and P_t are known. Assume also that the conditional distribution of α_t given Y_t is $\mathrm{N}(a_{t|t}, P_{t|t})$. The distribution of α_{t+1} given Y_t is $\mathrm{N}(a_{t+1}, P_{t+1})$. Our object is to calculate $a_{t|t}$, $P_{t|t}$, a_{t+1} and P_{t+1} when y_t is brought in. We refer to $a_{t|t}$ as the *filtered estimator* of the state α_t and a_{t+1} as the *one-step ahead predictor* of α_{t+1}. Their respective associated variances are $P_{t|t}$ and P_{t+1}.

An important part is played by the one-step ahead prediction error v_t of y_t. Then $v_t = y_t - a_t$ for $t = 1, \ldots, n$, and

$$\begin{aligned}
\mathrm{E}(v_t|Y_{t-1}) &= \mathrm{E}(\alpha_t + \varepsilon_t - a_t|Y_{t-1}) = a_t - a_t = 0, \\
\mathrm{Var}(v_t|Y_{t-1}) &= \mathrm{Var}(\alpha_t + \varepsilon_t - a_t|Y_{t-1}) = P_t + \sigma_\varepsilon^2, \\
\mathrm{E}(v_t|\alpha_t, Y_{t-1}) &= \mathrm{E}(\alpha_t + \varepsilon_t - a_t|\alpha_t, Y_{t-1}) = \alpha_t - a_t, \\
\mathrm{Var}(v_t|\alpha_t, Y_{t-1}) &= \mathrm{Var}(\alpha_t + \varepsilon_t - a_t|\alpha_t, Y_{t-1}) = \sigma_\varepsilon^2,
\end{aligned} \tag{2.7}$$

for $t = 2, \ldots, n$. When Y_t is fixed, Y_{t-1} and y_t are fixed so Y_{t-1} and v_t are fixed and vice versa. Consequently, $p(\alpha_t|Y_t) = p(\alpha_t|Y_{t-1}, v_t)$. We have

$$\begin{aligned}
p(\alpha_t|Y_{t-1}, v_t) &= p(\alpha_t, v_t|Y_{t-1})/p(v_t|Y_{t-1}) \\
&= p(\alpha_t|Y_{t-1})p(v_t|\alpha_t, Y_{t-1})/p(v_t|Y_{t-1}) \\
&= \text{constant} \times \exp(-\frac{1}{2}Q),
\end{aligned} \tag{2.8}$$

where

$$\begin{aligned} Q &= (\alpha_t - a_t)^2/P_t + (v_t - \alpha_t + a_t)^2/\sigma_\varepsilon^2 - v_t^2/(P_t + \sigma_\varepsilon^2) \\ &= \left(\frac{1}{P_t} + \frac{1}{\sigma_\varepsilon^2}\right)(\alpha_t - a_t)^2 - 2(\alpha_t - a_t)\frac{v_t}{\sigma_\varepsilon^2} + \left(\frac{1}{\sigma_\varepsilon^2} - \frac{1}{P_t + \sigma_\varepsilon^2}\right)v_t^2 \\ &= \frac{P_t + \sigma_\varepsilon^2}{P_t\,\sigma_\varepsilon^2}\left(\alpha_t - a_t - \frac{P_t\,v_t}{P_t + \sigma_\varepsilon^2}\right)^2. \end{aligned} \qquad (2.9)$$

Thus

$$p(\alpha_t|Y_t) = \mathrm{N}\left(a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}v_t\ ,\ \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2}\right). \qquad (2.10)$$

But $a_{t|t}$ and $P_{t|t}$ have been defined such that $p(\alpha_t|Y_t) = \mathrm{N}(a_{t|t}, P_{t|t})$. It follows that

$$a_{t|t} = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}v_t, \qquad (2.11)$$

$$P_{t|t} = \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2}. \qquad (2.12)$$

Since $a_{t+1} = \mathrm{E}(\alpha_{t+1}|Y_t) = \mathrm{E}(\alpha_t + \eta_t|Y_t)$ and $P_{t+1} = \mathrm{Var}(\alpha_{t+1}|Y_t) = \mathrm{Var}(\alpha_t + \eta_t|Y_t)$ from (2.3), we have

$$\begin{aligned} a_{t+1} &= \mathrm{E}(\alpha_t|Y_t) \ = \ a_{t|t}, \\ P_{t+1} &= \mathrm{Var}(\alpha_t|Y_t) + \sigma_\eta^2 \ = \ P_{t|t} + \sigma_\eta^2, \end{aligned}$$

giving

$$a_{t+1} = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}v_t, \qquad (2.13)$$

$$P_{t+1} = \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2} + \sigma_\eta^2, \qquad (2.14)$$

for $t = 2, \ldots, n$. For $t = 1$ we delete the symbol Y_{t-1} in the above derivation and we find that all results from (2.7) to (2.13) hold for $t = 1$ as well as for $t = 2, \ldots, n$.

In order to make these results consistent with the treatment of filtering for the general linear state space model in Subsection 4.3.1, we introduce the notation

$$F_t = \mathrm{Var}(v_t|Y_{t-1}) = P_t + \sigma_\varepsilon^2, \qquad K_t = P_t/F_t,$$

where F_t is referred to as the variance of the prediction error v_t and K_t is known as the *Kalman gain*. Using (2.11) to (2.14) we can then write the full set of relations for updating from time t to time $t + 1$ in the form

$$\begin{aligned} v_t &= y_t - a_t, & F_t &= P_t + \sigma_\varepsilon^2, \\ a_{t|t} &= a_t + K_t v_t, & P_{t|t} &= P_t(1 - K_t), \\ a_{t+1} &= a_t + K_t v_t, & P_{t+1} &= P_t(1 - K_t) + \sigma_\eta^2, \end{aligned} \tag{2.15}$$

for $t = 1, \dots, n$, where $K_t = P_t \,/\, F_t$.

We have assumed that a_1 and P_1 are known; however, more general initial specifications for a_1 and P_1 will be dealt with in Section 2.9. Relations (2.15) constitute the celebrated *Kalman filter* for the local level model. It should be noted that P_t depends only on σ_ε^2 and σ_η^2 and does not depend on Y_{t-1}. We include the case $t = n$ in (2.15) for convenience even though a_{n+1} and P_{n+1} are not normally needed for anything except forecasting. A set of relations such as (2.15) which enables us to calculate quantities for $t + 1$ given those for t is called a *recursion*.

2.2.2 Regression lemma

The above derivation of the Kalman filter can be regarded as an application of a regression lemma for the bivariate normal distribution. Suppose that x and y are jointly normally distributed variables with

$$\mathrm{E}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \qquad \mathrm{Var}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix},$$

with means μ_x and μ_y, variances σ_x^2 and σ_y^2, and covariance σ_{xy}. The joint distribution is

$$p(x, y) = p(y)\, p(x|y),$$

by the definition of the conditional density $p(x|y)$. But it can also be verified by direct multiplication. We have

$$p(x, y) = \frac{A}{2\pi} \exp\left\{ -\frac{1}{2}\sigma_y^{-2}(y - \mu_y)^2 - \frac{1}{2}\sigma_x^{-2}\left[x - \mu_x - \sigma_{xy}\sigma_y^{-2}(y - \mu_y)\right]^2 \right\},$$

where $A = \sigma_x^2 - \sigma_y^{-2}\sigma_{xy}$. It follows that the conditional distribution of x given y is normal and independent of y with mean and variance given by

$$\mathrm{E}(x|y) = \mu_x + \frac{\sigma_{xy}}{\sigma_y^2}(y - \mu_y), \qquad \mathrm{Var}(x|y) = \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2}.$$

To apply this lemma to the Kalman filter, let $v_t = y_t - a_t$ and keep Y_{t-1} fixed. Take $x = \alpha_t$ so that $\mu_x = a_t$ and $y = v_t$. It follows that $\mu_y = \mathrm{E}(v_t) = 0$. Then, $\sigma_x^2 = \mathrm{Var}(\alpha_t) = P_t$, $\sigma_y^2 = \mathrm{Var}(v_t) = \mathrm{Var}(\alpha_t - a_t + \varepsilon_t) = P_t + \sigma_\varepsilon^2$ and $\sigma_{xy} = P_t$. We obtain the conditional distribution for α_t given v_t by

$$\mathrm{E}(\alpha_t|v_t) = a_{t|t} = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \qquad \mathrm{Var}(\alpha_t|v_t) = P_{t|t} = \frac{P_t}{P_t + \sigma_\varepsilon^2}.$$

In a similar way we can obtain the equations for a_{t+1} and P_{t+1} by application of this regression lemma.

2.2.3 Bayesian treatment

To analyse the local level model from a Bayesian standpoint, we assume that the data are generated by model (2.3). In this approach α_t and y_t are regarded as a parameter and a constant, respectively. Before the observation y_t is taken, the prior distribution of α_t is $p(\alpha_t|Y_{t-1})$. The likelihood of α_t is $p(y_t|\alpha_t, Y_{t-1})$. The posterior distribution of α_t given y_t is given by the Bayes theorem which is proportional to the product of these. In particular we have

$$p(\alpha_t|Y_{t-1}, y_t) = p(\alpha_t|Y_{t-1})\, p(y_t|\alpha_t, Y_{t-1}) \,/\, p(y_t|Y_{t-1}).$$

Since $y_t = \alpha_t + \varepsilon_t$ we have $\mathrm{E}(y_t|Y_{t-1}) = a_t$ and $\mathrm{Var}(y_t|Y_{t-1}) = P_t + \sigma_\varepsilon^2$, so

$$p(\alpha_t|Y_{t-1}, y_t) = \text{constant} \times \exp(-\frac{1}{2}Q),$$

where

$$\begin{aligned} Q &= (\alpha_t - a_t)^2/P_t \;+\; (\alpha_t - a_t)^2/\sigma_\varepsilon^2 \;-\; (y_t - a_t)^2/(P_t + \sigma_\varepsilon^2) \\ &= \frac{P_t + \sigma_\varepsilon^2}{P_t\,\sigma_\varepsilon^2}\left(\alpha_t - a_t - \frac{P_t}{P_t + \sigma_\epsilon^2}(y_t - a_t)\right)^2. \end{aligned} \tag{2.16}$$

This is a normal density which we denote by $\mathrm{N}(a_{t|t}, P_{t|t})$. Thus the posterior mean and variance are

$$\begin{aligned} a_{t|t} &= a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \\ P_{t|t} &= \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2}, \end{aligned} \tag{2.17}$$

which are the same as (2.11) and (2.12) on putting $v_t = y_t - a_t$. The case $t = 1$ has the same form. Similarly, the posterior density of α_{t+1} given y_t is $p(\alpha_{t+1}|Y_{t-1}, y_t) = p(\alpha_t + \eta_t|Y_{t-1}, y_t)$, which is normal with mean $a_{t|t}$ and variance $P_{t|t} + \sigma_\eta^2$. Denoting this by $\mathrm{N}(a_{t+1}, P_{t+1})$, we have

$$\begin{aligned} a_{t+1} &= a_{t|t} = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \\ P_{t+1} &= P_{t|t} + \sigma_\eta^2 = \frac{P_t\,\sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2} + \sigma_\eta^2, \end{aligned} \tag{2.18}$$

which are, of course, the same as (2.13) and (2.14). It follows that the Kalman filter from a Bayesian point of view has the same form (2.15) as the Kalman filter from the standpoint of classical inference. This is an important result; as will be seen in Chapter 4 and later chapters, many inference results for the state α_t are the same whether approached from a classical or a Bayesian standpoint.

2.2.4 Minimum variance linear unbiased treatment

In some situations, some workers object to the assumption of normality in model (2.3) on the grounds that the observed time series they are concerned with do not appear to behave in a way that corresponds with the normal distribution. In these circumstances an alternative approach is to treat the filtering problem as a problem in the estimation of α_t and α_{t+1} given Y_t and to confine attention to estimates that are linear unbiased functions of y_t; we then choose those estimates that have minimum variance. We call these estimates *minimum variance linear unbiased estimates* (MVLUE).

Taking first the case of α_t, we seek an estimate $\bar{\alpha}_t$ which has the linear form $\bar{\alpha}_t = \beta + \gamma y_t$ where β and γ are constants given Y_{t-1} and which is unbiased in the sense that the estimation error $\bar{\alpha}_t - \alpha_t$ has zero mean in the conditional joint distribution of α_t and y_t given Y_{t-1}. We therefore have

$$\begin{aligned} \mathrm{E}(\bar{\alpha}_t - \alpha_t | Y_{t-1}) &= \mathrm{E}(\beta + \gamma y_t - \alpha_t | Y_{t-1}) \\ &= \beta + \gamma a_t - a_t = 0, \end{aligned} \tag{2.19}$$

so $\beta = a_t(1-\gamma)$ which gives $\bar{\alpha}_t = a_t + \gamma(y_t - a_t)$. Thus $\bar{\alpha}_t - \alpha_t = \gamma(\alpha_t - a_t + \varepsilon_t) - (\alpha_t - a_t)$. Now $\mathrm{Cov}(\alpha_t - a_t + \varepsilon_t, \alpha_t - a_t) = P_t$ so we have

$$\begin{aligned} \mathrm{Var}(\bar{\alpha}_t - \alpha_t | Y_{t-1}) &= \gamma^2 (P_t + \sigma_\varepsilon^2) - 2\gamma P_t + P_t \\ &= \left(P_t + \sigma_\varepsilon^2\right) \left(\gamma - \frac{P_t}{P_t + \sigma_\varepsilon^2}\right)^2 + P_t - \frac{P_t^2}{P_t + \sigma_\varepsilon^2}. \end{aligned} \tag{2.20}$$

This is minimised when $\gamma = P_t/(P_t + \sigma_\varepsilon^2)$ which gives

$$\bar{\alpha}_t = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \tag{2.21}$$

$$\mathrm{Var}(\bar{\alpha}_t - \alpha_t | Y_{t-1}) = \frac{P_t \, \sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2}. \tag{2.22}$$

Similarly, if we estimate α_{t+1} given Y_{t-1} by the linear function $\bar{\alpha}_{t+1}^* = \beta^* + \gamma^* y_t$ and require this to have the unbiasedness property $\mathrm{E}(\bar{\alpha}_{t+1}^* - \alpha_{t+1} | Y_{t-1}) = 0$, we find that $\beta^* = a_t(1-\gamma^*)$ so $\bar{\alpha}_{t+1}^* = a_t + \gamma^*(y_t - a_t)$. By the same argument as for $\bar{\alpha}_t$ we find that $\mathrm{Var}(\bar{\alpha}_{t+1}^* - \alpha_{t+1} | Y_{t-1})$ is minimised when $\gamma^* = P_t/(P_t + \sigma_\varepsilon^2)$ giving

$$\bar{\alpha}_{t+1}^* = a_t + \frac{P_t}{P_t + \sigma_\varepsilon^2}(y_t - a_t), \tag{2.23}$$

$$\mathrm{Var}(\bar{\alpha}_{t+1}^* - \alpha_{t+1} | Y_{t-1}) = \frac{P_t \, \sigma_\varepsilon^2}{P_t + \sigma_\varepsilon^2} + \sigma_\eta^2. \tag{2.24}$$

We have therefore shown that the estimates of $\bar{\alpha}_t$ and $\bar{\alpha}_{t+1}$ given by the MVLUE approach and their variances are exactly the same as the values $a_{t|t}$, a_{t+1}, $P_{t|t}$ and P_{t+1} in (2.11) to (2.14) that are obtained by assuming normality, both from a classical and from a Bayesian standpoint. It follows that the values given by the Kalman filter recursion (2.15) are MVLUE. We shall show in Subsection 4.3.1 that the same is true for the general linear Gaussian state space model (4.12).

2.2.5 Illustration

In this subsection we shall illustrate the output of the Kalman filter using observations from the river Nile. The data set consists of a series of readings of the annual flow volume at Aswan from 1871 to 1970. The series has been analysed by Cobb (1978) and Balke (1993). We analyse the data using the local level model (2.3) with $a_1 = 0$, $P_1 = 10^7$, $\sigma_\varepsilon^2 = 15,099$ and $\sigma_\eta^2 = 1,469.1$. The values for a_1 and P_1 were chosen arbitrarily for illustrative purposes. The values for σ_ε^2 and σ_η^2 are the maximum likelihood estimates which we obtain in Subsection 2.10.3. The values of a_t together with the raw data, P_t, v_t and F_t, for $t = 2, \ldots, n$, given by the Kalman filter, are presented graphically in Fig. 2.1.

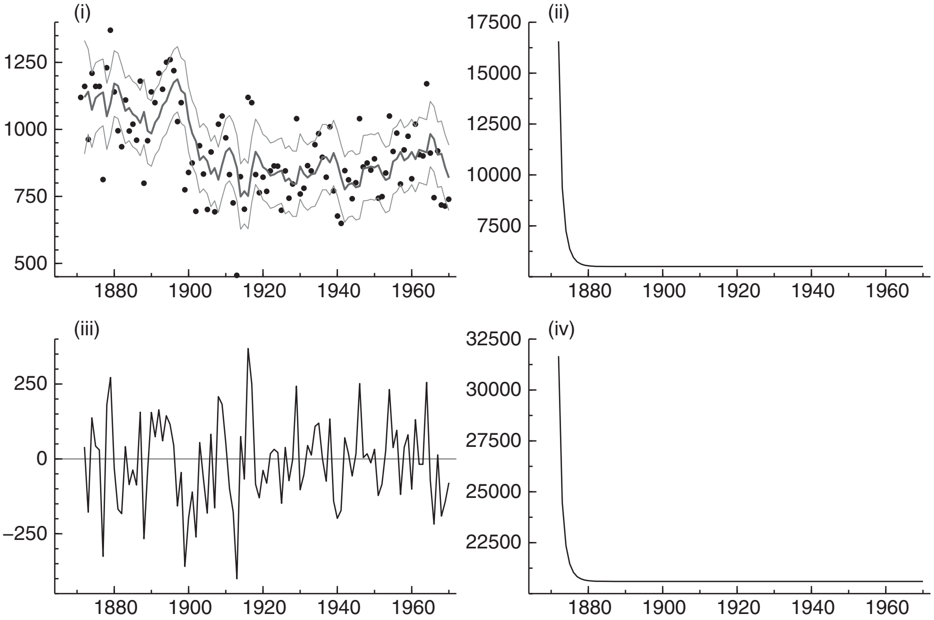

Fig. 2.1 Nile data and output of Kalman filter: (i) data (dots), filtered state a_t (solid line) and its 90% confidence intervals (light solid lines); (ii) filtered state variance P_t; (iii) prediction errors v_t; (iv) prediction variance F_t.

The most obvious feature of the four graphs is that P_t and F_t converge rapidly to constant values which confirms that the local level model has a steady state solution; for discussion of the concept of a steady state see Section 2.11. However, it was found that the fitted local level model converged numerically to a steady state in around 25 updates of P_t although the graph of P_t seems to suggest that the steady state was obtained after around 10 updates.

2.3 Forecast errors

The Kalman filter residual $v_t = y_t - a_t$ and its variance F_t are the one-step ahead forecast error and the one-step ahead forecast error variance of y_t given Y_{t-1} as defined in Section 2.2. The forecast errors $v_1, \dots, v_n$ are sometimes called *innovations* because they represent the new part of y_t that cannot be predicted from the past for $t = 1, \dots, n$. We shall make use of v_t and F_t for a variety of results in the next sections. It is therefore important to study them in detail.

2.3.1 Cholesky decomposition

First we show that $v_1, \dots, v_n$ are mutually independent. The joint density of $y_1, \dots, y_n$ is

$$p(y_1, \dots, y_n) = p(y_1) \prod_{t=2}^{n} p(y_t | Y_{t-1}). \tag{2.25}$$

We then transform from $y_1, \dots, y_n$ to $v_1, \dots, v_n$. Since each v_t equals y_t minus a linear function of $y_1, \dots, y_{t-1}$ for $t = 2, \dots, n$, the Jacobian is one. From (2.25) and making the substitution we have

$$p(v_1, \dots, v_n) = \prod_{t=1}^{n} p(v_t), \tag{2.26}$$

since $p(v_1) = p(y_1)$ and $p(v_t) = p(y_t | Y_{t-1})$ for $t = 2, \dots, n$. Consequently, the v_t's are independently distributed.

We next show that the forecast errors v_t are effectively obtained from a Cholesky decomposition of the observation vector Y_n. The Kalman filter recursions compute the forecast error v_t as a linear function of the initial mean a_1 and the observations $y_1, \dots, y_t$ since

$$\begin{aligned}
v_1 &= y_1 - a_1, \\
v_2 &= y_2 - a_1 - K_1(y_1 - a_1), \\
v_3 &= y_3 - a_1 - K_2(y_2 - a_1) - K_1(1 - K_2)(y_1 - a_1), \quad \text{and so on.}
\end{aligned}$$

It should be noted that K_t does not depend on the initial mean a_1 and the observations $y_1, \dots, y_n$; it depends only on the initial state variance P_1 and the disturbance variances σ_ε^2 and σ_η^2. Using the definitions in (2.4), we have

$$v = C(Y_n - \mathbf{1}a_1), \quad \text{with} \quad v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

where matrix C is the lower triangular matrix

$$C = \begin{bmatrix} 1 & 0 & 0 & & 0 \\ c_{21} & 1 & 0 & & 0 \\ c_{31} & c_{32} & 1 & & 0 \\ & & & \ddots & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & 1 \end{bmatrix},$$

$$\begin{aligned} c_{i,i-1} &= -K_{i-1}, \\ c_{ij} &= -(1-K_{i-1})(1-K_{i-2})\cdots(1-K_{j+1})K_j, \end{aligned} \tag{2.27}$$

for $i = 2, \ldots, n$ and $j = 1, \ldots, i-2$. The distribution of v is therefore

$$v \sim \mathrm{N}(0, C\Omega C'), \tag{2.28}$$

where $\Omega = \mathrm{Var}(Y_n)$ as given by (2.4). On the other hand we know from (2.7), (2.15) and (2.26) that $\mathrm{E}(v_t) = 0$, $\mathrm{Var}(v_t) = F_t$ and $\mathrm{Cov}(v_t, v_j) = 0$, for $t, j = 1, \ldots, n$ and $t \neq j$; therefore,

$$v \sim \mathrm{N}(0, F), \quad \text{with} \quad F = \begin{bmatrix} F_1 & 0 & 0 & & 0 \\ 0 & F_2 & 0 & & 0 \\ 0 & 0 & F_3 & & 0 \\ & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & F_n \end{bmatrix}.$$

It follows that $C\Omega C' = F$. The transformation of a symmetric positive definite matrix (say Ω) into a diagonal matrix (say F) using a lower triangular matrix (say C) by means of the relation $C\Omega C' = F$ is known as the *Cholesky decomposition* of the symmetric matrix. The Kalman filter can therefore be regarded as essentially a Cholesky decomposition of the variance matrix implied by the local level model (2.3). This result is important for understanding the role of the Kalman filter and it will be used further in Subsections 2.5.4 and 2.10.1. Note also that $F^{-1} = (C')^{-1}\Omega^{-1}C^{-1}$ so we have $\Omega^{-1} = C'F^{-1}C$.

2.3.2 Error recursions

Define the *state estimation error* as

$$x_t = \alpha_t - a_t, \quad \text{with} \quad \mathrm{Var}(x_t) = P_t. \tag{2.29}$$

We now show that the state estimation errors x_t and forecast errors v_t are linear functions of the initial state error x_1 and the disturbances ε_t and η_t analogously to the way that α_t and y_t are linear functions of the initial state and the disturbances for $t = 1, \dots, n$. It follows directly from the Kalman filter relations (2.15) that

$$\begin{aligned} v_t &= y_t - a_t \\ &= \alpha_t + \varepsilon_t - a_t \\ &= x_t + \varepsilon_t, \end{aligned}$$

and

$$\begin{aligned} x_{t+1} &= \alpha_{t+1} - a_{t+1} \\ &= \alpha_t + \eta_t - a_t - K_t v_t \\ &= x_t + \eta_t - K_t(x_t + \varepsilon_t) \\ &= L_t x_t + \eta_t - K_t \varepsilon_t, \end{aligned}$$

where

$$L_t = 1 - K_t = \sigma_\varepsilon^2 / F_t. \tag{2.30}$$

Thus analogously to the local level model relations

$$y_t = \alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = \alpha_t + \eta_t,$$

we have the error relations

$$v_t = x_t + \varepsilon_t, \qquad x_{t+1} = L_t x_t + \eta_t - K_t \varepsilon_t, \qquad t = 1, \dots, n, \tag{2.31}$$

with $x_1 = \alpha_1 - a_1$. These relations will be used in the next section. We note that P_t, F_t, K_t and L_t do not depend on the initial state mean a_1 or the observations $y_1, \dots, y_n$ but only on the initial state variance P_1 and the disturbance variances σ_ε^2 and σ_η^2. We note also that the recursion for P_{t+1} in (2.15) can alternatively be derived by

$$\begin{aligned} P_{t+1} &= \mathrm{Var}(x_{t+1}) = \mathrm{Cov}(x_{t+1}, \alpha_{t+1}) = \mathrm{Cov}(x_{t+1}, \alpha_t + \eta_t) \\ &= L_t \mathrm{Cov}(x_t, \alpha_t + \eta_t) + \mathrm{Cov}(\eta_t, \alpha_t + \eta_t) - K_t \mathrm{Cov}(\varepsilon_t, \alpha_t + \eta_t) \\ &= L_t P_t + \sigma_\eta^2 \ = \ P_t(1 - K_t) + \sigma_\eta^2. \end{aligned}$$

2.4 State smoothing

2.4.1 Smoothed state

We now consider the estimation of $\alpha_1, \dots, \alpha_n$ in model (2.3) given the entire sample Y_n. Since all distributions are normal, the conditional density of α_t given Y_n

is $\mathrm{N}(\hat{\alpha}_t, V_t)$ where $\hat{\alpha}_t = \mathrm{E}(\alpha_t|Y_n)$ and $V_t = \mathrm{Var}(\alpha_t|Y_n)$. We call $\hat{\alpha}_t$ the *smoothed state*, V_t the *smoothed state variance* and the operation of calculating $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ *state smoothing*. Similar arguments to those in Subsections 2.2.3 and 2.2.4 can be used to justify the same formulae for a Bayesian analysis and a MVLUE approach.

The forecast errors $v_1, \ldots, v_n$ are mutually independent and $v_t, \ldots, v_n$ are independent of $y_1, \ldots, y_{t-1}$ with zero means. Moreover, when $y_1, \ldots, y_n$ are fixed, Y_{t-1} and $v_t, \ldots, v_n$ are fixed and vice versa. By an extension of the lemma of Subsection 2.2.2 to the multivariate case we have the regression relation for the conditional distribution of α_t and $v_t, \ldots, v_n$ given Y_{t-1},

$$\hat{\alpha}_t = a_t + \sum_{j=t}^{n} \mathrm{Cov}(\alpha_t, v_j) F_j^{-1} v_j. \tag{2.32}$$

Now $\mathrm{Cov}(\alpha_t, v_j) = \mathrm{Cov}(x_t, v_j)$ for $j = t, \ldots, n$, and

$$\begin{aligned}\mathrm{Cov}(x_t, v_t) &= \mathrm{E}[x_t(x_t + \varepsilon_t)] = \mathrm{Var}(x_t) = P_t,\\ \mathrm{Cov}(x_t, v_{t+1}) &= \mathrm{E}[x_t(x_{t+1} + \varepsilon_{t+1})] = \mathrm{E}[x_t(L_t x_t + \eta_t - K_t \varepsilon_t)] = P_t L_t,\end{aligned}$$

where x_t is defined in (2.29) and L_t in (2.30). Similarly,

$$\begin{aligned}\mathrm{Cov}(x_t, v_{t+2}) &= P_t L_t L_{t+1},\\ &\vdots\\ \mathrm{Cov}(x_t, v_n) &= P_t L_t L_{t+1} \ldots L_{n-1}.\end{aligned} \tag{2.33}$$

Substituting in (2.32) gives

$$\begin{aligned}\hat{\alpha}_t &= a_t + P_t \frac{v_t}{F_t} + P_t L_t \frac{v_{t+1}}{F_{t+1}} + P_t L_t L_{t+1} \frac{v_{t+2}}{F_{t+2}} + \ldots + P_t L_t L_{t+1} \ldots L_{n-1} \frac{v_n}{F_n}\\ &= a_t + P_t r_{t-1},\end{aligned}$$

where

$$\begin{aligned}r_{t-1} = \frac{v_t}{F_t} &+ L_t \frac{v_{t+1}}{F_{t+1}} + L_t L_{t+1} \frac{v_{t+2}}{F_{t+2}} + L_t L_{t+1} L_{t+2} \frac{v_{t+3}}{F_{t+3}} + \ldots +\\ &+ L_t L_{t+1} \ldots L_{n-1} \frac{v_n}{F_n}\end{aligned} \tag{2.34}$$

is a weighted sum of innovations after $t-1$. The value of this at time t is

$$\begin{aligned}r_t = \frac{v_{t+1}}{F_{t+1}} &+ L_{t+1} \frac{v_{t+2}}{F_{t+2}} + L_{t+1} L_{t+2} \frac{v_{t+3}}{F_{t+3}} + \cdots\\ &+ L_{t+1} L_{t+2} \ldots L_{n-1} \frac{v_n}{F_n}.\end{aligned} \tag{2.35}$$

Obviously, $r_n = 0$ since no observations are available after time n. By substituting from (2.35) into (2.34), it follows that the values of r_{t-1} can be evaluated using the backwards recursion

$$r_{t-1} = \frac{v_t}{F_t} + L_t r_t, \tag{2.36}$$

with $r_n = 0$, for $t = n, n-1, \ldots, 1$. The smoothed state can therefore be calculated by the backwards recursion

$$r_{t-1} = F_t^{-1} v_t + L_t r_t, \qquad \hat{\alpha}_t = a_t + P_t r_{t-1}, \qquad t = n, \ldots, 1, \tag{2.37}$$

with $r_n = 0$. The relations in (2.37) are collectively called the *state smoothing recursion.*

2.4.2 Smoothed state variance

The error variance of the smoothed state, $V_t = \mathrm{Var}(\alpha_t|Y_n)$, is derived in a similar way. By the multivariate extension of the regression lemma of Subsection 2.2.2 applied to the conditional distribution of α_t and $v_t, \ldots, v_n$ given Y_{t-1} we have

$$\begin{aligned} V_t &= \mathrm{Var}(\alpha_t|Y_n) = \mathrm{Var}(\alpha_t|Y_{t-1}, v_t, \ldots, v_n) \\ &= P_t - \sum_{j=t}^{n} [\mathrm{Cov}(\alpha_t, v_j)]^2 F_j^{-1}, \end{aligned} \tag{2.38}$$

where the expressions for $\mathrm{Cov}(\alpha_t, v_j)$ $\mathrm{Cov}(x_t, v_j)$ are given by (2.33). Substituting these into (2.38) leads to

$$\begin{aligned} V_t &= P_t - P_t^2 \frac{1}{F_t} - P_t^2 L_t^2 \frac{1}{F_{t+1}} - P_t^2 L_t^2 L_{t+1}^2 \frac{1}{F_{t+2}} - \cdots - P_t^2 L_t^2 L_{t+1}^2 \cdots L_{n-1}^2 \frac{1}{F_n} \\ &= P_t - P_t^2 N_{t-1}, \end{aligned} \tag{2.39}$$

where

$$\begin{aligned} N_{t-1} = {} & \frac{1}{F_t} + L_t^2 \frac{1}{F_{t+1}} + L_t^2 L_{t+1}^2 \frac{1}{F_{t+2}} + L_t^2 L_{t+1}^2 L_{t+2}^2 \frac{1}{F_{t+3}} + \cdots \\ & + L_t^2 L_{t+1}^2 \cdots L_{n-1}^2 \frac{1}{F_n}, \end{aligned} \tag{2.40}$$

is a weighted sum of the inverse variances of innovations after time $t-1$. Its value at time t is

$$N_t = \frac{1}{F_{t+1}} + L_{t+1}^2 \frac{1}{F_{t+2}} + L_{t+1}^2 L_{t+2}^2 \frac{1}{F_{t+3}} + \cdots + L_{t+1}^2 L_{t+2}^2 \cdots L_{n-1}^2 \frac{1}{F_n}, \tag{2.41}$$

and, obviously, $N_n = 0$ since no variances are available after time n. Substituting from (2.41) into (2.40) it follows that the value for N_{t-1} can be calculated using the backwards recursion

$$N_{t-1} = \frac{1}{F_t} + L_t^2 N_t, \tag{2.42}$$

with $N_n = 0$, for $t = n, n-1, \ldots, 1$. We observe from (2.35) and (2.41) that $N_t = \mathrm{Var}(r_t)$ since the forecast errors v_t are independent.

By combining these results, the error variance of the smoothed state can be calculated by the backwards recursion

$$N_{t-1} = F_t^{-1} + L_t^2 N_t, \qquad V_t = P_t - P_t^2 N_{t-1}, \qquad t = n, \ldots, 1, \tag{2.43}$$

with $N_n = 0$. The relations in (2.43) are collectively called the *state variance smoothing recursion*. From the standard error $\sqrt{V_t}$ of $\hat{\alpha}_t$ we can construct confidence intervals for α_t for $t = 1, \ldots, n$. It is also possible to derive the smoothed covariances between the states, that is, $\mathrm{Cov}(\alpha_t, \alpha_s | Y_n)$, $t \neq s$, using similar arguments. We shall not give them here but will derive them for the general case in Section 4.7.

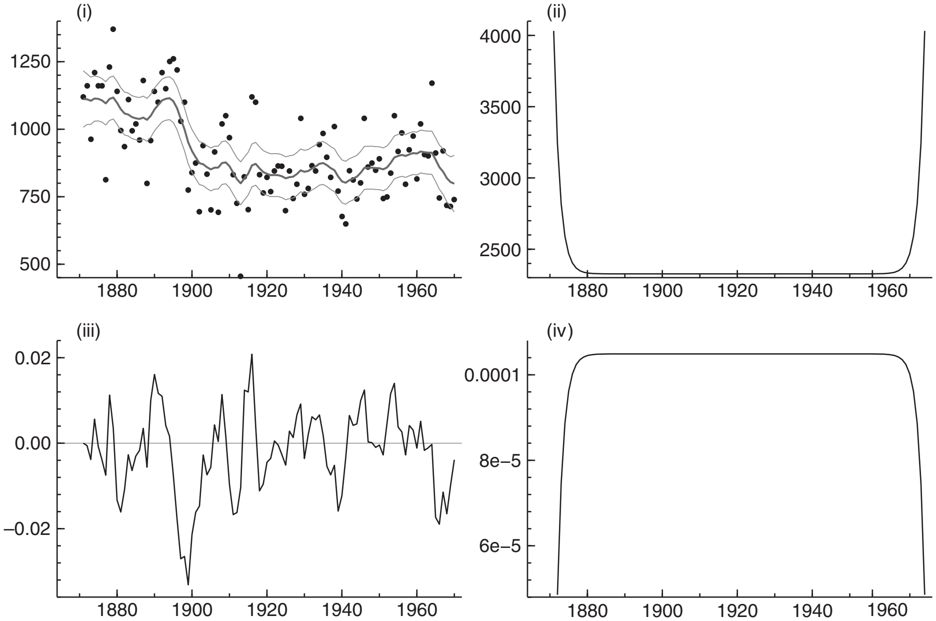

Fig. 2.2 Nile data and output of state smoothing recursion: (i) data (dots), smoothed state $\hat{\alpha}_t$ and its 90% confidence intervals; (ii) smoothed state variance V_t; (iii) smoothing cumulant r_t; (iv) smoothing variance cumulant N_t.

2.4.3 Illustration

We now show the results of state smoothing for the Nile data of Subsection 2.2.5 using the same local level model. The Kalman filter is applied first and the output v_t, F_t, a_t and P_t is stored for $t = 1, \dots, n$. Figure 2.2 presents the output of the backwards smoothing recursions (2.37) and (2.43); that is $\hat{\alpha}_t$, V_t, r_t and N_t. The plot of $\hat{\alpha}_t$ includes the 90% confidence bands for α_t. The graph of $\mathrm{Var}(\alpha_t|Y_n)$ shows that the conditional variance of α_t is larger at the beginning and end of the sample, as it obviously should be on intuitive grounds. Comparing the graphs of a_t and $\hat{\alpha}_t$ in Fig. 2.1 and 2.2, we see that the graph of $\hat{\alpha}_t$ is much smoother than that of a_t, except at time points close to the end of the series, as it should be.

2.5 Disturbance smoothing

In this section we consider the calculation of the smoothed observation disturbance $\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t|Y_n) = y_t - \hat{\alpha}_t$ and the smoothed state disturbance $\hat{\eta}_t = \mathrm{E}(\eta_t|Y_n) = \hat{\alpha}_{t+1} - \hat{\alpha}_t$ together with their error variances. Of course, these could be calculated directly from a knowledge of $\hat{\alpha}_1, \dots, \hat{\alpha}_n$ and covariances $\mathrm{Cov}(\alpha_t, \alpha_j|Y_n)$ for $j \leq t$. However, it turns out to be computationally advantageous to compute them from r_t and N_t without first calculating $\hat{\alpha}_t$, particularly for the general model discussed in Chapter 4. The merits of smoothed disturbances are discussed in Section 4.5. For example, the estimates $\hat{\varepsilon}_t$ and $\hat{\eta}_t$ are useful for detecting outliers and structural breaks, respectively; see Subsection 2.12.2. For the sake of brevity we shall restrict the treatment of this section to classical inference based on the assumption of normality as in model (2.3).

In order to economise on the amount of algebra in this chapter we shall present the required recursions for the local level model without proof, referring the reader to Section 4.5 for derivations of the analogous recursions for the general model.

2.5.1 Smoothed observation disturbances

From (4.58) in Section 4.5.1, the smoothed observation disturbance $\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t|Y_n)$ is calculated by

$$\hat{\varepsilon}_t = \sigma_\varepsilon^2 u_t, \qquad t = n, \dots, 1, \tag{2.44}$$

where

$$u_t = F_t^{-1} v_t - K_t r_t, \tag{2.45}$$

and where the recursion for r_t is given by (2.36). The scalar u_t is referred to as the *smoothing error*. Similarly, from (4.65) in Section 4.5.2, the smoothed variance $\mathrm{Var}(\varepsilon_t|Y_n)$ is obtained by

$$\mathrm{Var}(\varepsilon_t|Y_n) = \sigma_\varepsilon^2 - \sigma_\varepsilon^4 D_t, \qquad t = n, \dots, 1, \tag{2.46}$$

where

$$D_t = F_t^{-1} + K_t^2 N_t, \tag{2.47}$$

and where the recursion for N_t is given by (2.42). Since from (2.35) v_t is independent of r_t, and $\mathrm{Var}(r_t) = N_t$, we have

$$\mathrm{Var}(u_t) = \mathrm{Var}\big(F_t^{-1}v_t - K_t r_t\big) = F_t^{-2}\,\mathrm{Var}(v_t) + K_t^2\,\mathrm{Var}(r_t) = D_t.$$

Consequently, from (2.44) we obtain $\mathrm{Var}(\hat{\varepsilon}_t) = \sigma_\varepsilon^4 D_t$.

Note that the methods for calculating $\hat{\alpha}_t$ and $\hat{\varepsilon}_t$ are consistent since $K_t = P_t F_t^{-1}$, $L_t = 1 - K_t = \sigma_\varepsilon^2 F_t^{-1}$ and

$$\begin{aligned}
\hat{\varepsilon}_t &= y_t - \hat{\alpha}_t \\
&= y_t - a_t - P_t r_{t-1} \\
&= v_t - P_t\big(F_t^{-1}v_t + L_t r_t\big) \\
&= F_t^{-1}v_t(F_t - P_t) - \sigma_\varepsilon^2 P_t F_t^{-1} r_t \\
&= \sigma_\varepsilon^2\big(F_t^{-1}v_t - K_t r_t\big), \qquad t = n, \ldots, 1.
\end{aligned}$$

Similar equivalences can be shown for V_t and $\mathrm{Var}(\varepsilon_t|Y_n)$.

2.5.2 Smoothed state disturbances

From (4.63) in Subsection 4.5.1, the smoothed mean of the disturbance $\hat{\eta}_t = \mathrm{E}(\eta_t|Y_n)$ is calculated by

$$\hat{\eta}_t = \sigma_\eta^2 r_t, \qquad t = n, \ldots, 1, \tag{2.48}$$

where the recursion for r_t is given by (2.36). Similarly, from (4.68) in Subsection 4.5.2, the smoothed variance $\mathrm{Var}(\eta_t|Y_n)$ is computed by

$$\mathrm{Var}(\eta_t|Y_n) = \sigma_\eta^2 - \sigma_\eta^4 N_t, \qquad t = n, \ldots, 1, \tag{2.49}$$

where the recursion for N_t is given by (2.42). Since $\mathrm{Var}(r_t) = N_t$, we have $\mathrm{Var}(\hat{\eta}_t) = \sigma_\eta^4 N_t$. These results are interesting because they give an interpretation to the values r_t and N_t; they are the scaled smoothed estimator of $\eta_t = \alpha_{t+1} - \alpha_t$ and its unconditional variance, respectively.

The method of calculating $\hat{\eta}_t$ is consistent with the definition $\eta_t = \alpha_{t+1} - \alpha_t$ since

$$\begin{aligned}
\hat{\eta}_t &= \hat{\alpha}_{t+1} - \hat{\alpha}_t \\
&= a_{t+1} + P_{t+1} r_t - a_t - P_t r_{t-1} \\
&= a_t + K_t v_t - a_t + P_t L_t r_t + \sigma_\eta^2 r_t - P_t\big(F_t^{-1}v_t + L_t r_t\big) \\
&= \sigma_\eta^2 r_t.
\end{aligned}$$

Similar consistencies can be shown for N_t and $\mathrm{Var}(\eta_t|Y_n)$.

2.5.3 Illustration

The smoothed disturbances and their related variances for the analysis of the Nile data and the local level model introduced in Subsection 2.2.5 are calculated by the above recursions and presented in Fig. 2.3. We note from the graphs of $\mathrm{Var}(\varepsilon_t|Y_n)$ and $\mathrm{Var}(\eta_t|Y_n)$ the extent that these conditional variances are larger at the beginning and end of the sample. Obviously, the plot of r_t in Fig. 2.2 and the plot of $\hat{\eta}_t$ in Fig. 2.3 are the same apart from a different scale.

2.5.4 Cholesky decomposition and smoothing

We now consider the calculation of $\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t|Y_n)$ by direct regression of $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ on the observation vector Y_n defined in (2.4) to obtain $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)'$, that is,

$$\begin{aligned}\hat{\varepsilon} &= \mathrm{E}(\varepsilon) + \mathrm{Cov}(\varepsilon, Y_n)\,\mathrm{Var}(Y_n)^{-1}[Y_n - \mathrm{E}(Y_n)] \\ &= \mathrm{Cov}(\varepsilon, Y_n)\Omega^{-1}(Y_n - 1a_1),\end{aligned}$$

where, here and later, when necessary, we treat Y_n as the observation vector $(y_1, \dots, y_n)'$. It is obvious from (2.6) that $\mathrm{Cov}(\varepsilon, Y_n) = \sigma_\varepsilon^2 I_n$; also, from the

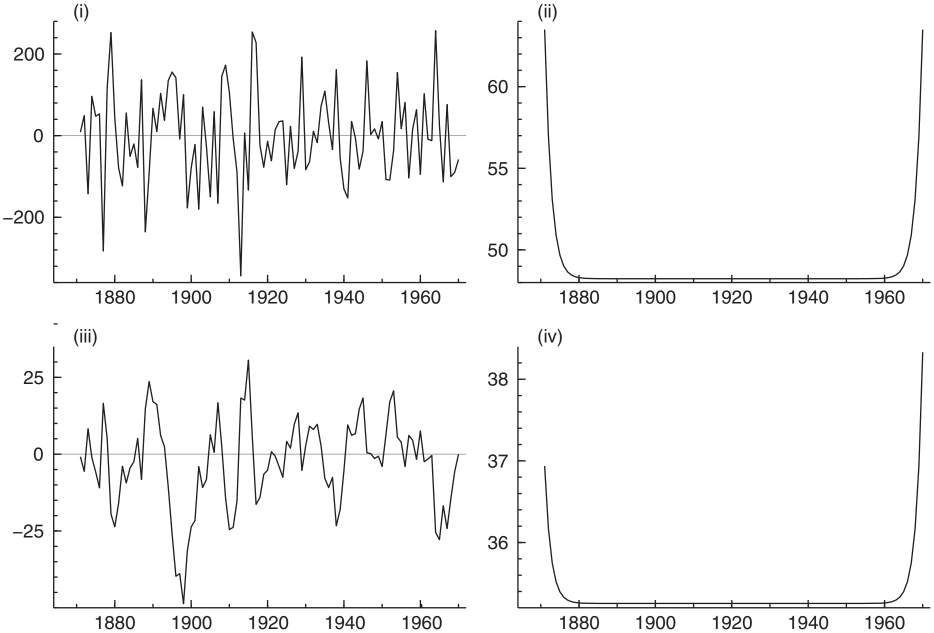

Fig. 2.3 Output of disturbance smoothing recursion: (i) observation error $\hat{\varepsilon}_t$; (ii) observation error variance $\mathrm{Var}(\varepsilon_t|Y_n)$; (iii) state error $\hat{\eta}_t$; (iv) state error variance $\mathrm{Var}(\eta_t|Y_n)$.

Cholesky decomposition considered in Subsection 2.3.1 we have $\Omega^{-1} = C'F^{-1}C$ and $C(Y_n - 1a_1) = v$. We therefore have

$$\hat{\varepsilon} = \sigma_\varepsilon^2 C'F^{-1}v,$$

which, by consulting the definitions of the lower triangular elements of C in (2.27), also leads to the disturbance equations (2.44) and (2.45). Thus

$$\hat{\varepsilon} = \sigma_\varepsilon^2 u, \qquad u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

where

$$u = C'F^{-1}v \quad \text{with} \quad v = C(Y_n - 1a_1).$$

It follows that

$$u = C'F^{-1}C(Y_n - 1a_1) = \Omega^{-1}(Y_n - 1a_1), \tag{2.50}$$

where $\Omega = \text{Var}(Y_n)$ and $F = C\Omega C'$, as is consistent with standard regression theory.

2.6 Simulation

It is simple to draw samples generated by the local level model (2.3). We first draw the random normal deviates

$$\varepsilon_t^+ \sim \text{N}\big(0, \sigma_\varepsilon^2\big), \qquad \eta_t^+ \sim \text{N}\big(0, \sigma_\eta^2\big), \qquad t = 1, \ldots, n. \tag{2.51}$$

Then we generate observations using the local level recursion as follows

$$y_t^+ = \alpha_t^+ + \varepsilon_t^+, \qquad \alpha_{t+1}^+ = \alpha_t^+ + \eta_t^+, \qquad t = 1, \ldots, n, \tag{2.52}$$

for some starting value α_1^+.

For the implementation of classical and Bayesian simulation methods and for the treatment of nonlinear and non-Gaussian models, which will be discussed in Part II of this book, we may require samples generated by the local level model conditional on the observed time series $y_1, \ldots, y_n$. Such samples can be obtained by use of the simulation smoother developed for the general linear Gaussian state space model in Section 4.9. For the local level model, a simulated sample for the disturbances ε_t, $t = 1, \ldots, n$, given the observations $y_1, \ldots, y_n$ can be obtained using the method of mean corrections as discussed in Subsection 4.9.1. It requires the drawing of the samples ε_t^+ and η_t^+ as in (2.51) and using them to draw y_t^+ as in (2.52). Then, a conditional draw for ε_t given Y_n is obtained by

$$\tilde{\varepsilon}_t = \varepsilon_t^+ - \hat{\varepsilon}_t^+ + \hat{\varepsilon}_t, \tag{2.53}$$

for $t = 1, \ldots, n$, where $\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t|Y_n)$ and $\hat{\varepsilon}_t^+ = \mathrm{E}(\varepsilon_t|Y_n^+)$ with $Y_n^+ = (y_1^+, \ldots, y_n^+)'$ and where both are computed via the disturbance smoothing equations (2.44) and (2.45). This set of computations is sufficient to obtain a conditional draw of ε_t given Y_n, for $t = 1, \ldots, n$. Given a sample $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$, we obtain simulated samples for α_t and η_t via the relations

$$\tilde{\alpha}_t = y_t - \tilde{\varepsilon}_t, \qquad \tilde{\eta}_t = \tilde{\alpha}_{t+1} - \tilde{\alpha}_t,$$

for $t = 1, \ldots, n$.

2.6.1 Illustration

To illustrate the difference between simulating a sample from the local level model unconditionally and simulating a sample conditional on the observations, we consider the Nile data and the local level model of Subsection 2.2.5. In Fig. 2.4 (i) we present the smoothed state $\hat{\alpha}_t$ and a sample generated by the local level model unconditionally. The two series have seemingly nothing in common. In the next panel, again the smoothed state is presented but now together with a sample

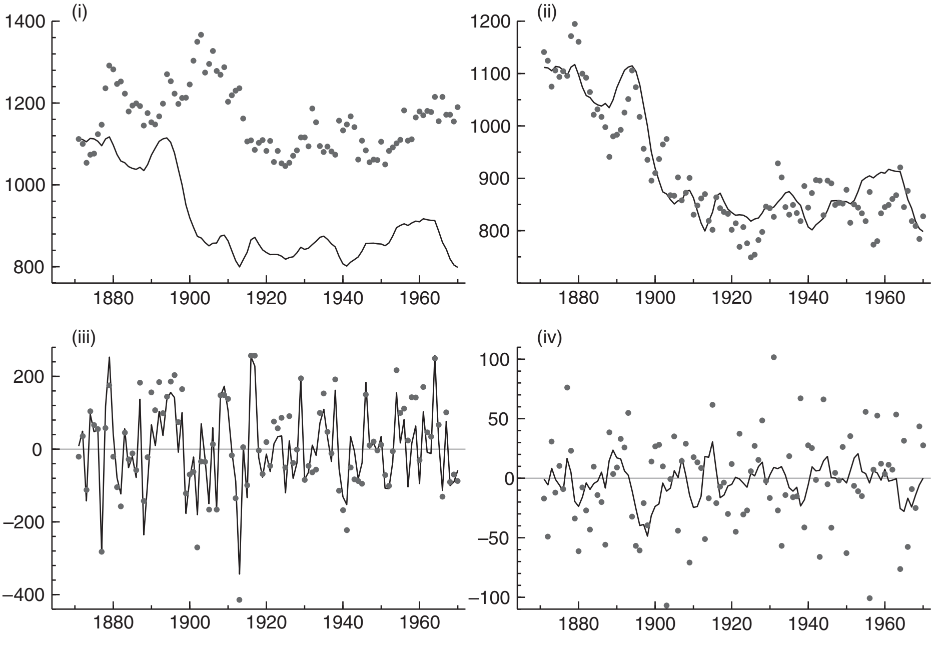

Fig. 2.4 Simulation: (i) smoothed state $\hat{\alpha}_t$ (solid line) and sample α_t^+ (dots); (ii) smoothed state $\hat{\alpha}_t$ (solid line) and sample $\tilde{\alpha}_t$ (dots); (iii) smoothed observation error $\hat{\varepsilon}_t$ (solid line) and sample $\tilde{\varepsilon}_t$ (dots); (iv) smoothed state error $\hat{\eta}_t$ (solid line) and sample $\tilde{\eta}_t$ (dots).

generated conditional on the observations. Here we see that the generated sample is much closer to $\hat{\alpha}_t$. The remaining two panels present the smoothed disturbances together with a sample from the corresponding disturbances conditional on the observations.

2.7 Missing observations

A considerable advantage of the state space approach is the ease with which missing observations can be dealt with. Suppose we have a local level model where observations y_j, with $j = \tau, \dots, \tau^* - 1$, are missing for $1 < \tau < \tau^* \leq n$. For the filtering stage, the most obvious way to deal with the situation is to define a new series y_t^* where $y_t^* = y_t$ for $t = 1, \dots, \tau - 1$ and $y_t^* = y_{t+\tau^*-\tau}$ for $t = \tau, \dots, n^*$ with $n^* = n - (\tau^* - \tau)$. The model for y_t^* with time scale $t = 1, \dots, n^*$ is then the same as (2.3) with $y_t = y_t^*$ except that $\alpha_\tau = \alpha_{\tau-1} + \eta_{\tau-1}$ where $\eta_{\tau-1} \sim \mathrm{N}[0, (\tau^* - \tau)\sigma_\eta^2]$. Filtering for this model can be treated by the methods developed in Chapter 4 for the general state space model. The treatment is readily extended if more than one group of observations is missing.

It is, however, easier and more transparent to proceed as follows, using the original time domain. For filtering at times $t = \tau, \dots, \tau^* - 1$, we have

$$\begin{aligned}
\mathrm{E}(\alpha_t|Y_t) = \mathrm{E}(\alpha_t|Y_{\tau-1}) &= \mathrm{E}\left(\alpha_\tau + \sum_{j=\tau}^{t-1} \eta_j \middle| Y_{\tau-1}\right) = a_\tau, \\
\mathrm{E}(\alpha_{t+1}|Y_t) = \mathrm{E}(\alpha_{t+1}|Y_{\tau-1}) &= \mathrm{E}\left(\alpha_\tau + \sum_{j=\tau}^{t} \eta_j \middle| Y_{\tau-1}\right) = a_\tau, \\
\mathrm{Var}(\alpha_t|Y_t) = \mathrm{Var}(\alpha_t|Y_{\tau-1}) &= \mathrm{Var}\left(\alpha_\tau + \sum_{j=\tau}^{t-1} \eta_j \middle| Y_{\tau-1}\right) = P_\tau + (t-\tau)\sigma_\eta^2, \\
\mathrm{Var}(\alpha_{t+1}|Y_t) = \mathrm{Var}(\alpha_t|Y_{\tau-1}) &= \mathrm{Var}\left(\alpha_\tau + \sum_{j=\tau}^{t} \eta_j \middle| Y_{\tau-1}\right) = P_\tau + (t-\tau+1)\sigma_\eta^2.
\end{aligned}$$

We can compute them recursively by

$$\begin{aligned}
a_{t|t} &= a_t, & P_{t|t} &= P_t, \\
a_{t+1} &= a_t, & P_{t+1} &= P_t + \sigma_\eta^2,
\end{aligned} \qquad t = \tau, \dots, \tau^* - 1, \tag{2.54}$$

the remaining values a_t and P_t being given as before by (2.15) for $t = 1, \dots, \tau - 1$ and $t = \tau^*, \dots, n$. The consequence is that we can use the original filter (2.15) for all t by taking $K_t = 0$ at the missing time points. The same procedure is used when more than one group of observations is missing. It follows that allowing for missing observations when using the Kalman filter is extremely simple.

The forecast error recursions from which we derive the smoothing recursions are given by (2.31). These error-updating equations at the missing time points become

$$v_t = x_t + \varepsilon_t, \qquad x_{t+1} = x_t + \eta_t, \qquad t = \tau, \ldots, \tau^* - 1,$$

since $K_t = 0$ and therefore $L_t = 1$. The covariances between the state at the missing time points and the innovations after the missing period are given by

$$\begin{aligned} \mathrm{Cov}(\alpha_t, v_{\tau^*}) &= P_t, \\ \mathrm{Cov}(\alpha_t, v_j) &= P_t L_{\tau^*} L_{\tau^*+1} \ldots L_{j-1}, \quad j = \tau^* + 1, \ldots, n, \quad t = \tau, \ldots, \tau^* - 1. \end{aligned}$$

By deleting the terms associated with the missing time points, the state smoothing equation (2.32) for the missing time points becomes

$$\hat{\alpha}_t = a_t + \sum_{j=\tau^*}^{n} \mathrm{Cov}(\alpha_t, v_j) F_j^{-1} v_j, \qquad t = \tau, \ldots, \tau^* - 1.$$

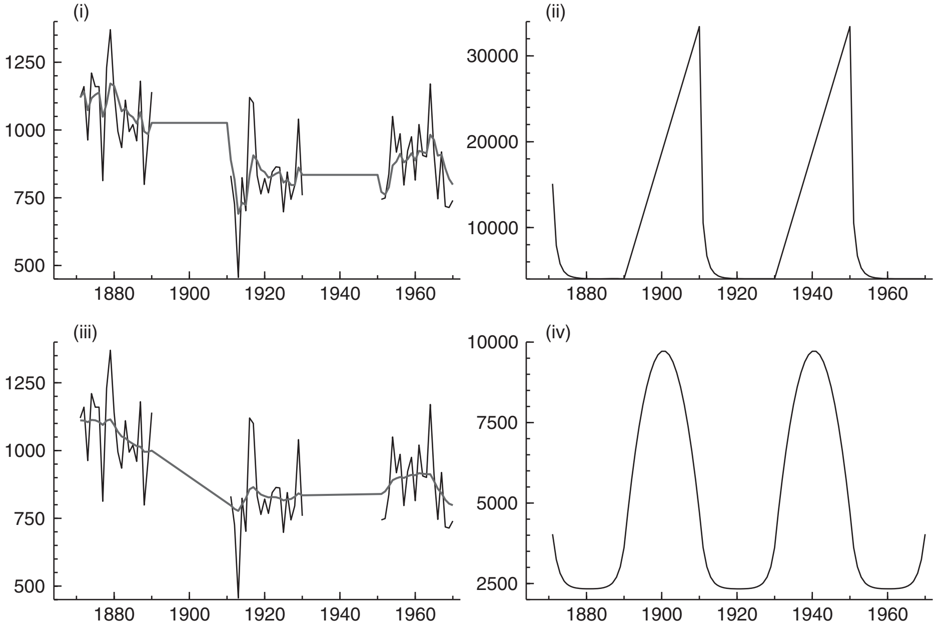

Fig. 2.5 Filtering and smoothing output when observations are missing: (i) data and filtered state a_t (extrapolation); (ii) filtered state variance P_t; (iii) data and smoothed state $\hat{\alpha}_t$ (interpolation); (iv) smoothed state variance V_t.

Substituting the covariance terms into this and taking into account the definition (2.34) leads directly to

$$r_{t-1} = r_t, \qquad \hat{\alpha}_t = a_t + P_t r_{t-1}, \qquad t = \tau, \dots, \tau^* - 1. \tag{2.55}$$

The consequence is that we can use the original state smoother (2.37) for all t by taking $K_t = 0$, and hence $L_t = 1$, at the missing time points. This device applies to any missing observation within the sample period. In the same way the equations for the variance of the state error and the smoothed disturbances can be obtained by putting $K_t = 0$ at missing time points.

2.7.1 Illustration

Here we consider the Nile data and the same local level model as before; however, we treat the observations at time points $21, \dots, 40$ and $61, \dots, 80$ as missing. The Kalman filter is applied first and the output v_t, F_t, a_t and P_t is stored for $t = 1, \dots, n$. Then, the state smoothing recursions are applied. The first two graphs in Fig. 2.5 are the Kalman filter values of a_t and P_t, respectively. The last two graphs are the smoothing output $\hat{\alpha}_t$ and V_t, respectively.

Note that the application of the Kalman filter to missing observations can be regarded as extrapolation of the series to the missing time points, while smoothing at these points is effectively interpolation.

2.8 Forecasting

Let $\bar{y}_{n+j}$ be the minimum mean square error forecast of y_{n+j} given the time series $y_1, \dots, y_n$ for $j = 1, 2, \dots, J$ with J as some pre-defined positive integer. By minimum mean square error forecast here we mean the function $\bar{y}_{n+j}$ of $y_1, \dots, y_n$ which minimises $\mathrm{E}[(y_{n+j} - \bar{y}_{n+j})^2 | Y_n]$. Then $\bar{y}_{n+j} = \mathrm{E}(y_{n+j} | Y_n)$. This follows immediately from the well-known result that if x is a random variable with mean μ the value of λ that minimises $\mathrm{E}(x - \lambda)^2$ is $\lambda = \mu$; see Exercise 4.14.3. The variance of the forecast error is denoted by $\bar{F}_{n+j} = \mathrm{Var}(y_{n+j} | Y_n)$. The theory of forecasting for the local level model turns out to be surprisingly simple; we merely regard forecasting as filtering the observations $y_1, \dots, y_n, y_{n+1}, \dots, y_{n+J}$ using the recursion (2.15) and treating the last J observations $y_{n+1}, \dots, y_{n+J}$ as missing, that is, taking $K_t = 0$ in (2.15).

Letting $\bar{a}_{n+j} = \mathrm{E}(\alpha_{n+j} | Y_n)$ and $\bar{P}_{n+j} = \mathrm{Var}(\alpha_{n+j} | Y_n)$, it follows immediately from equation (2.54) with $\tau = n + 1$ and $\tau^* = n + J$ in §2.7 that

$$\bar{a}_{n+j+1} = \bar{a}_{n+j}, \qquad \bar{P}_{n+j+1} = \bar{P}_{n+j} + \sigma_\eta^2, \qquad j = 1, \dots, J - 1,$$

with $\bar{a}_{n+1} = a_{n+1}$ and $\bar{P}_{n+1} = P_{n+1}$ obtained from the Kalman filter (2.15). Furthermore, we have

$$\bar{y}_{n+j} = \mathrm{E}(y_{n+j}|Y_n) = \mathrm{E}(\alpha_{n+j}|Y_n) + \mathrm{E}(\varepsilon_{n+j}|Y_n) = \bar{a}_{n+j},$$
$$\bar{F}_{n+j} = \mathrm{Var}(y_{n+j}|Y_n) = \mathrm{Var}(\alpha_{n+j}|Y_n) + \mathrm{Var}(\varepsilon_{n+j}|Y_n) = \bar{P}_{n+j} + \sigma_\varepsilon^2,$$

for $j = 1, \dots, J$. The consequence is that the Kalman filter can be applied for $t = 1, \dots, n + J$ where we treat the observations at times $n + 1, \dots, n + J$ as missing. Thus we conclude that forecasts and their error variances are delivered by applying the Kalman filter in a routine way with $K_t = 0$ for $t = n{+}1, \dots, n{+}J$. The same property holds for the general linear Gaussian state space model as we shall show in Section 4.11. For a Bayesian treatment a similar argument can be used to show that the posterior mean and variance of the forecast of y_{n+j} is obtained by treating $y_{n+1}, \dots, y_{n+j}$ as missing values, for $j = 1, \dots, J$.

2.8.1 Illustration

The Nile data set is now extended by 30 missing observations allowing the computation of forecasts for the observations $y_{101}, \dots, y_{130}$. Only the Kalman filter

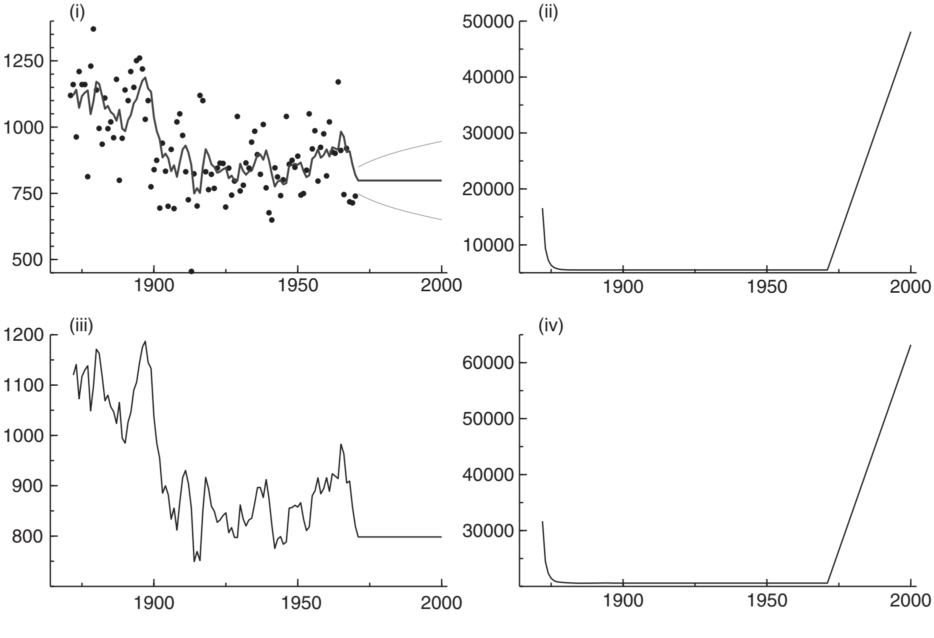

Fig. 2.6 Nile data and output of forecasting: (i) data (dots), state forecast a_t and 50% confidence intervals; (ii) state variance P_t; (iii) observation forecast $\mathrm{E}(y_t|Y_{t-1})$; (iv) observation forecast variance F_t.

is required. The graphs in Fig. 2.6 contain $\hat{y}_{n+j|n} = a_{n+j|n}$, $P_{n+j|n}$, $a_{n+j|n}$ and $F_{n+j|n}$, respectively, for $j = 1, \dots, J$ with $J = 30$. The confidence interval for $\mathrm{E}(y_{n+j}|Y_n)$ is $\hat{y}_{n+j|n} \pm k\sqrt{F}_{n+j|n}$ where k is determined by the required probability of inclusion; in Fig. 2.6 this probability is 50%.

2.9 Initialisation

We assumed in our treatment of the linear Gaussian model in previous sections that the distribution of the initial state α_1 is $\mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known. We now consider how to start up the filter (2.15) when nothing is known about the distribution of α_1, which is the usual situation in practice. In this situation it is reasonable to represent α_1 as having a *diffuse prior* density, that is, fix a_1 at an arbitrary value and let $P_1 \to \infty$. From (2.15) we have

$$v_1 = y_1 - a_1, \qquad F_1 = P_1 + \sigma_\varepsilon^2,$$

and, by substituting into the equations for a_2 and P_2 in (2.15), it follows that

$$a_2 = a_1 + \frac{P_1}{P_1 + \sigma_\varepsilon^2}(y_1 - a_1), \tag{2.56}$$

$$\begin{aligned} P_2 &= P_1\left(1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2}\right) + \sigma_\eta^2 \\ &= \frac{P_1}{P_1 + \sigma_\varepsilon^2}\sigma_\varepsilon^2 + \sigma_\eta^2. \end{aligned} \tag{2.57}$$

Letting $P_1 \to \infty$, we obtain $a_2 = y_1$, $P_2 = \sigma_\varepsilon^2 + \sigma_\eta^2$; we can then proceed normally with the Kalman filter (2.15) for $t = 2, \dots, n$. This process is called *diffuse initialisation* of the Kalman filter and the resulting filter is called *the diffuse Kalman filter*. We note the interesting fact that the same values of a_t and P_t for $t = 2, \dots, n$ can be obtained by treating y_1 as fixed and taking $\alpha_1 \sim \mathrm{N}(y_1, \sigma_\varepsilon^2)$. Specifically, we have $a_{1|1} = y_1$ and $P_{1|1} = \sigma_\varepsilon^2$. It follows from (2.18) for $t = 1$ that $a_2 = y_1$ and $P_2 = \sigma_\varepsilon^2 + \sigma_\eta^2$. This is intuitively reasonable in the absence of information about the marginal distribution of α_1 since $(y_1 - \alpha_1) \sim \mathrm{N}(0, \sigma_\varepsilon^2)$.

We also need to take account of the diffuse distribution of the initial state α_1 in the smoothing recursions. It is shown above that the filtering equations for $t = 2, \dots, n$ are not affected by letting $P_1 \to \infty$. Therefore, the state and disturbance smoothing equations are also not affected for $t = n, \dots, 2$ since these only depend on the Kalman filter output. From (2.37), the smoothed mean of the state α_1 is given by

$$\begin{aligned} \hat{\alpha}_1 &= a_1 + P_1\left[\frac{1}{P_1 + \sigma_\varepsilon^2}v_1 + \left(1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2}\right)r_1\right] \\ &= a_1 + \frac{P_1}{P_1 + \sigma_\varepsilon^2}v_1 + \frac{P_1}{P_1 + \sigma_\varepsilon^2}\sigma_\varepsilon^2 r_1. \end{aligned}$$

Letting $P_1 \to \infty$, we obtain $\hat{\alpha}_1 = a_1 + v_1 + \sigma_\varepsilon^2 r_1$ and by substituting for v_1 we have

$$\hat{\alpha}_1 = y_1 + \sigma_\varepsilon^2 r_1.$$

The smoothed conditional variance of the state α_1 given Y_n is, from (2.43)

$$\begin{aligned} V_1 &= P_1 - P_1^2 \left[\frac{1}{P_1 + \sigma_\varepsilon^2} + \left(1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2} \right)^2 N_1 \right] \\ &= P_1 \left(1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2} \right) - \left(\frac{P_1}{P_1 + \sigma_\varepsilon^2} \right)^2 \sigma_\varepsilon^4 N_1 \\ &= \left(\frac{P_1}{P_1 + \sigma_\varepsilon^2} \right) \sigma_\varepsilon^2 - \left(\frac{P_1}{P_1 + \sigma_\varepsilon^2} \right)^2 \sigma_\varepsilon^4 N_1. \end{aligned}$$

Letting $P_1 \to \infty$, we obtain $V_1 = \sigma_\varepsilon^2 - \sigma_\varepsilon^4 N_1$.

The smoothed means of the disturbances for $t = 1$ are given by

$$\hat{\varepsilon}_1 = \sigma_\varepsilon^2 u_1, \quad \text{with} \quad u_1 = \frac{1}{P_1 + \sigma_\varepsilon^2} v_1 - \frac{P_1}{P_1 + \sigma_\varepsilon^2} r_1,$$

and $\hat{\eta}_1 = \sigma_\eta^2 r_1$. Letting $P_1 \to \infty$, we obtain $\hat{\varepsilon}_1 = -\sigma_\varepsilon^2 r_1$. Note that r_1 depends on the Kalman filter output for $t = 2, \ldots, n$. The smoothed variances of the disturbances for $t = 1$ depend on D_1 and N_1 of which only D_1 is affected by $P_1 \to \infty$; using (2.47),

$$D_1 = \frac{1}{P_1 + \sigma_\varepsilon^2} + \left(\frac{P_1}{P_1 + \sigma_\varepsilon^2} \right)^2 N_1.$$

Letting $P_1 \to \infty$, we obtain $D_1 = N_1$ and therefore $\mathrm{Var}(\hat{\varepsilon}_1) = \sigma_\varepsilon^4 N_1$. The variance of the smoothed estimate of η_1 remains unaltered as $\mathrm{Var}(\hat{\eta}_1) = \sigma_\eta^4 N_1$.

The initial smoothed state $\hat{\alpha}_1$ under diffuse conditions can also be obtained by assuming that y_1 is fixed and $\alpha_1 = y_1 - \varepsilon_1$ where $\varepsilon_1 \sim N(0, \sigma_\varepsilon^2)$. For example, for the smoothed mean of the state at $t = 1$, we have now only $n - 1$ varying y_t's so that

$$\hat{\alpha}_1 = a_1 + \sum_{j=2}^{n} \frac{\mathrm{Cov}(\alpha_1, v_j)}{F_j} v_j$$

with $a_1 = y_1$. It follows from (2.56) that $a_2 = a_1 = y_1$. Further, $v_2 = y_2 - a_2 = \alpha_2 + \varepsilon_2 - y_1 = \alpha_1 + \eta_1 + \varepsilon_2 - y_1 = -\varepsilon_1 + \eta_1 + \varepsilon_2$. Consequently, $\mathrm{Cov}(\alpha_1, v_2) = \mathrm{Cov}(-\varepsilon_1, -\varepsilon_1 + \eta_1 + \varepsilon_2) = \sigma_\varepsilon^2$. We therefore have from (2.32),

$$\begin{aligned} \hat{\alpha}_1 &= a_1 + \frac{\sigma_\varepsilon^2}{F_2} v_2 + \frac{(1 - K_2)\sigma_\varepsilon^2}{F_3} v_3 + \frac{(1 - K_2)(1 - K_3)\sigma_\varepsilon^2}{F_4} v_4 + \cdots \\ &= y_1 + \sigma_\varepsilon^2 r_1, \end{aligned}$$

as before with r_1 as defined in (2.34) for $t = 1$. The equations for the remaining $\hat{\alpha}_t$'s are the same as previously. The same results may be obtained by Bayesian arguments.

Use of a diffuse prior for initialisation is the approach preferred by most time series analysts in the situation where nothing is known about the initial value α_1. However, some workers find the diffuse approach uncongenial because they regard the assumption of an infinite variance as unnatural since all observed time series have finite values. From this point of view an alternative approach is to assume that α_1 is an unknown constant to be estimated from the data by maximum likelihood. The simplest form of this idea is to estimate α_1 by maximum likelihood from the first observation y_1. Denote this maximum likelihood estimate by $\hat{\alpha}_1$ and its variance by $\mathrm{Var}(\hat{\alpha}_1)$. We then initialise the Kalman filter by taking $a_{1|1} = \hat{\alpha}_1$ and $P_{1|1} = \mathrm{Var}(\hat{\alpha}_1)$. Since when α_1 is fixed $y_1 \sim \mathrm{N}(\alpha_1, \sigma_\varepsilon^2)$, we have $\hat{\alpha}_1 = y_1$ and $\mathrm{Var}(\hat{\alpha}_1) = \sigma_\varepsilon^2$. We therefore initialise the filter by taking $a_{1|1} = y_1$ and $P_{1|1} = \sigma_\varepsilon^2$. But these are the same values as we obtain by assuming that α_1 is diffuse. It follows that we obtain the same initialisation of the Kalman filter by representing α_1 as a random variable with infinite variance as by assuming that it is fixed and unknown and estimating it from y_1. We shall show that a similar result holds for the general linear Gaussian state space model in Subsection 5.7.3.

2.10 Parameter estimation

We now consider the fitting of the local level model to data from the standpoint of classical inference. In effect, this amounts to deriving formulae on the assumption that the additional parameters are known and then replacing these by their maximum likelihood estimates. Bayesian treatments will be considered for the general linear Gaussian model in Chapter 13. Parameters in state space models are often called *hyperparameters*, possibly to distinguish them from elements of state vectors which can plausibly be thought of as random parameters; however, in this book we shall just call them *additional parameters*, since with the usual meaning of the word parameter this is what they are. We will discuss methods for calculating the loglikelihood function and the maximisation of it with respect to the additional parameters, σ_ε^2 and σ_η^2.

2.10.1 Loglikelihood evaluation

Since

$$p(y_1, \ldots, y_t) = p(Y_{t-1})p(y_t|Y_{t-1}),$$

for $t = 2, \ldots, n$, the joint density of $y_1, \ldots, y_n$ can be expressed as

$$p(Y_n) = \prod_{t=1}^{n} p(y_t|Y_{t-1}),$$

where $p(y_1|Y_0) = p(y_1)$. Now $p(y_t|Y_{t-1}) = \mathrm{N}(a_t, F_t)$ and $v_t = y_t - a_t$ so on taking logs and assuming that a_1 and P_1 are known the loglikelihood is given by

$$\log L = \log p(Y_n) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left(\log F_t + \frac{v_t^2}{F_t}\right). \tag{2.58}$$

The exact loglikelihood can therefore be constructed easily from the Kalman filter (2.15).

Alternatively, let us derive the loglikelihood for the local level model from the representation (2.4). This gives

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Omega| - \frac{1}{2}(Y_n - a_1\mathtt{1})'\Omega^{-1}(Y_n - a_1\mathtt{1}), \tag{2.59}$$

which follows from the multivariate normal distribution $Y_n \sim \mathrm{N}(a_1\mathtt{1}, \Omega)$. Using results from §2.3.1, $\Omega = CFC'$, $|C| = 1$, $\Omega^{-1} = C'F^{-1}C$ and $v = C(Y_n - a_1\mathtt{1})$; it follows that

$$\log|\Omega| = \log|CFC'| = \log|C||F||C| = \log|F|,$$

and

$$(Y_n - a_1\mathtt{1})'\Omega^{-1}(Y_n - a_1\mathtt{1}) = v'F^{-1}v.$$

Substitution and using the results $\log|F| = \sum_{t=1}^{n}\log F_t$ and $v'F^{-1}v = \sum_{t=1}^{n} F_t^{-1}v_t^2$ lead directly to (2.58).

The loglikelihood in the diffuse case is derived as follows. All terms in (2.58) remain finite as $P_1 \to \infty$ with Y_n fixed except the term for $t = 1$. It thus seems reasonable to remove the influence of P_1 as $P_1 \to \infty$ by defining the *diffuse loglikelihood* as

$$\begin{aligned}\log L_d &= \lim_{P_1\to\infty}\left(\log L + \frac{1}{2}\log P_1\right)\\ &= -\frac{1}{2}\lim_{P_1\to\infty}\left(\log\frac{F_1}{P_1} + \frac{v_1^2}{F_1}\right) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\left(\log F_t + \frac{v_t^2}{F_t}\right)\\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\left(\log F_t + \frac{v_t^2}{F_t}\right), \qquad (2.60)\end{aligned}$$

since $F_1/P_1 \to 1$ and $v_1^2/F_1 \to 0$ as $P_1 \to \infty$. Note that v_t and F_t remain finite as $P_1 \to \infty$ for $t = 2, \ldots, n$.

Since P_1 does not depend on σ_ε^2 and σ_η^2, the values of σ_ε^2 and σ_η^2 that maximise $\log L$ are identical to the values that maximise $\log L + \frac{1}{2}\log P_1$. As $P_1 \to \infty$, these latter values converge to the values that maximise $\log L_d$ because first and second derivatives with respect to σ_ε^2 and σ_η^2 converge, and second derivatives are finite and strictly negative. It follows that the maximum likelihood estimators of σ_ε^2 and σ_η^2 obtained by maximising (2.58) converge to the values obtained by maximising (2.60) as $P_1 \to \infty$.

We estimate the unknown parameters σ_ε^2 and σ_η^2 by maximising expression (2.58) or (2.60) numerically according to whether a_1 and P_1 are known or unknown. In practice it is more convenient to maximise numerically with respect to the quantities $\psi_\varepsilon = \log \sigma_\varepsilon^2$ and $\psi_\eta = \log \sigma_\eta^2$. An efficient algorithm for numerical maximisation is implemented in the *STAMP* 8.3 package of Koopman, Harvey, Doornik and Shephard (2010). This optimisation procedure is based on the quasi-Newton scheme BFGS for which details are given in Subsection 7.3.2.

2.10.2 Concentration of loglikelihood

It can be advantageous to re-parameterise the model prior to maximisation in order to reduce the dimensionality of the numerical search for the estimation of the parameters. For example, for the local level model we can put $q = \sigma_\eta^2/\sigma_\varepsilon^2$ to obtain the model

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t, \qquad & \varepsilon_t &\sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \\ \alpha_{t+1} &= \alpha_t + \eta_t, \qquad & \eta_t &\sim \mathrm{N}\big(0, q\sigma_\varepsilon^2\big), \end{aligned}$$

and estimate the pair σ_ε^2, q in preference to $\sigma_\varepsilon^2, \sigma_\eta^2$. Put $P_t^* = P_t/\sigma_\varepsilon^2$ and $F_t^* = F_t/\sigma_\varepsilon^2$; from (2.15) and Section 2.9, we have

$$\begin{aligned} v_t &= y_t - a_t, \qquad & F_t^* &= P_t^* + 1, \\ a_{t+1} &= a_t + K_t v_t, \qquad & P_{t+1}^* &= P_t^*(1 - K_t) + q, \end{aligned}$$

where $K_t = P_t/F_t = P_t^*/F_t^*$ for $t = 2, \dots, n$ and these relations are initialised with $a_2 = y_1$ and $P_2^* = 1 + q$. Note that F_t^* depends on q but not on σ_ε^2. The loglikelihood (2.60) then becomes

$$\log L_d = -\frac{n}{2}\log(2\pi) - \frac{n-1}{2}\log \sigma_\varepsilon^2 - \frac{1}{2}\sum_{t=2}^{n}\left(\log F_t^* + \frac{v_t^2}{\sigma_\varepsilon^2 F_t^*}\right). \tag{2.61}$$

By maximising (2.61) with respect to σ_ε^2, for given $F_2^*, \dots, F_n^*$, we obtain

$$\hat\sigma_\varepsilon^2 = \frac{1}{n-1}\sum_{t=2}^{n}\frac{v_t^2}{F_t^*}. \tag{2.62}$$

The value of $\log L_d$ obtained by substituting $\hat\sigma_\varepsilon^2$ for σ_ε^2 in (2.61) is called the *concentrated diffuse loglikelihood* and is denoted by $\log L_{dc}$, giving

$$\log L_{dc} = -\frac{n}{2}\log(2\pi) - \frac{n-1}{2} - \frac{n-1}{2}\log \hat\sigma_\varepsilon^2 - \frac{1}{2}\sum_{t=2}^{n}\log F_t^*. \tag{2.63}$$

This is maximised with respect to q by a one-dimensional numerical search.

Table 2.1 Estimation of parameters of local level model by maximum likelihood.

Iteration	q	ψ	Score	Loglikelihood
0	1	0	−3.32	−495.68
1	0.0360	−3.32	0.93	−492.53
2	0.0745	−2.60	0.25	−492.10
3	0.0974	−2.32	−0.001	−492.07
4	0.0973	−2.33	0.0	−492.07

2.10.3 Illustration

The estimates of the variances σ_ε^2 and $\sigma_\eta^2 = q\sigma_\varepsilon^2$ for the Nile data are obtained by maximising the concentrated diffuse loglikelihood (2.63) with respect to ψ where $q = \exp(\psi)$. In Table 2.1 the iterations of the BFGS procedure are reported starting with $\psi = 0$. The relative percentage change of the loglikelihood goes down very rapidly and convergence is achieved after 4 iterations. The final estimate for ψ is -2.33 and hence the estimate of q is $\hat{q} = 0.097$. The estimate of σ_ε^2 given by (2.62) is 15099 which implies that the estimate of σ_η^2 is $\hat{\sigma}_\eta^2 = \hat{q}\hat{\sigma}_\varepsilon^2 = 0.097 \times 15099 = 1469.1$.

2.11 Steady state

We now consider whether the Kalman filter (2.15) converges to a *steady state* as $n \to \infty$. This will be the case if P_t converges to a positive value, $\bar{P}$ say. Obviously, we would then have $F_t \to \bar{P} + \sigma_\varepsilon^2$ and $K_t \to \bar{P}/(\bar{P} + \sigma_\varepsilon^2)$. To check whether there is a steady state, put $P_{t+1} = P_t = \bar{P}$ in (2.15) and verify whether the resulting equation in $\bar{P}$ has a positive solution. The equation is

$$\bar{P} = \bar{P}\left(1 - \frac{\bar{P}}{\bar{P} + \sigma_\varepsilon^2}\right) + \sigma_\eta^2,$$

which reduces to the quadratic

$$x^2 - xq - q = 0, \tag{2.64}$$

where $x = \bar{P}/\sigma_\varepsilon^2$ and $q = \sigma_\eta^2/\sigma_\varepsilon^2$, with the solution

$$x = \left(q + \sqrt{q^2 + 4q}\right)/2.$$

This is positive when $q > 0$ which holds for nontrivial models. The other solution to (2.64) is inapplicable since it is negative for $q > 0$. Thus all non-trivial local level models have a steady state solution.

The practical advantage of knowing that a model has a steady state solution is that, after convergence of P_t to $\bar{P}$ has been verified as close enough, we can stop computing F_t and K_t and the filter (2.15) reduces to the single relation

$$a_{t+1} = a_t + \bar{K} v_t,$$

with $\bar{K} = \bar{P}/(\bar{P} + \sigma_\varepsilon^2)$ and $v_t = y_t - a_t$. While this has little consequence for the simple local level model we are concerned with here, it is a useful property for the more complicated models we shall consider in Chapter 4, where P_t can be a large matrix.

2.12 Diagnostic checking

2.12.1 Diagnostic tests for forecast errors

The assumptions underlying the local level model are that the disturbances ε_t and η_t are normally distributed and serially independent with constant variances. On these assumptions the standardised one-step ahead forecast errors

$$e_t = \frac{v_t}{\sqrt{F_t}}, \qquad t = 1, \ldots, n, \tag{2.65}$$

(or for $t = 2, \ldots, n$ in the diffuse case) are also normally distributed and serially independent with unit variance. We can check that these properties hold by means of the following large-sample diagnostic tests:

- Normality
  The first four moments of the standardised forecast errors are given by

$$m_1 = \frac{1}{n} \sum_{t=1}^{n} e_t,$$

$$m_q = \frac{1}{n} \sum_{t=1}^{n} (e_t - m_1)^q, \qquad q = 2, 3, 4,$$

  with obvious modifications in the diffuse case. Skewness and kurtosis are denoted by S and K, respectively, and are defined as

$$S = \frac{m_3}{\sqrt{m_2^3}}, \qquad K = \frac{m_4}{m_2^2},$$

  and it can be shown that when the model assumptions are valid they are asymptotically normally distributed as

$$S \sim \mathrm{N}\left(0, \frac{6}{n}\right), \qquad K \sim \mathrm{N}\left(3, \frac{24}{n}\right);$$

see Bowman and Shenton (1975). Standard statistical tests can be used to check whether the observed values of S and K are consistent with their asymptotic densities. They can also be combined as

$$N = n\left\{\frac{S^2}{6} + \frac{(K-3)^2}{24}\right\},$$

which asymptotically has a χ^2 distribution with 2 degrees of freedom on the null hypothesis that the normality assumption is valid. The *QQ plot* is a graphical display of ordered residuals against their theoretical quantiles. The 45 degree line is taken as a reference line (the closer the residual plot to this line, the better the match).

- Heteroscedasticity

  A simple test for heteroscedasticity is obtained by comparing the sum of squares of two exclusive subsets of the sample. For example, the statistic

$$H(h) = \frac{\sum_{t=n-h+1}^{n} e_t^2}{\sum_{t=1}^{h} e_t^2},$$

  is $F_{h,h}$-distributed for some preset positive integer h, under the null hypothesis of homoscedasticity. Here, e_t is defined in (2.65) and the sum of h squared forecast errors in the denominator starts at $t = 2$ in the diffuse case.

- Serial correlation

  When the local level model holds, the standardised forecast errors are serially uncorrelated as we have shown in Subsection 2.3.1. Therefore, the correlogram of the forecast errors should reveal serial correlation insignificant. A standard portmanteau test statistic for serial correlation is based on the Box–Ljung statistic suggested by Ljung and Box (1978). This is given by

$$Q(k) = n(n+2)\sum_{j=1}^{k}\frac{c_j^2}{n-j},$$

  for some preset positive integer k where c_j is the jth correlogram value

$$c_j = \frac{1}{nm_2}\sum_{t=j+1}^{n}(e_t - m_1)(e_{t-j} - m_1).$$

More details on diagnostic checking will be given in Section 7.5.

2.12.2 Detection of outliers and structural breaks

The standardised smoothed residuals are given by

$$\begin{aligned}
u_t^* &= \hat{\varepsilon}_t/\sqrt{\mathrm{Var}(\hat{\varepsilon}_t)} = D_t^{-\frac{1}{2}}u_t,\\
r_t^* &= \hat{\eta}_t/\sqrt{\mathrm{Var}(\hat{\eta}_t)} = N_t^{-\frac{1}{2}}r_t, \qquad t = 1,\ldots,n;
\end{aligned}$$

see Section 2.5 for details on computing the quantities u_t, D_t, r_t and N_t. Harvey and Koopman (1992) refer to these standardised residuals as *auxiliary residuals* and they investigate their properties in detail. For example, they show that the auxiliary residuals are autocorrelated and they discuss their autocorrelation function. The auxiliary residuals can be useful in detecting outliers and structural breaks in time series because $\hat{\varepsilon}_t$ and $\hat{\eta}_t$ are estimators of ε_t and η_t. An outlier in a series that we postulate as generated by the local level model is indicated by a large (positive or negative) value for $\hat{\varepsilon}_t$, or u_t^*, and a break in the level α_{t+1} is indicated by a large (positive or negative) value for $\hat{\eta}_t$, or r_t^*. A discussion of the use of auxiliary residuals for the general model will be given in Section 7.5.

2.12.3 Illustration

We consider the fitted local level model for the Nile data as obtained in Subsection 2.10.3. A plot of e_t is given in Fig. 2.7 together with the histogram, the QQ plot and the correlogram. These plots are satisfactory and they suggest that the assumptions underlying the local level model are valid for the Nile data. This is largely confirmed by the following diagnostic test statistics

$$S = -0.03, \quad K = 0.09, \quad N = 0.05, \quad H(33) = 0.61, \quad Q(9) = 8.84.$$

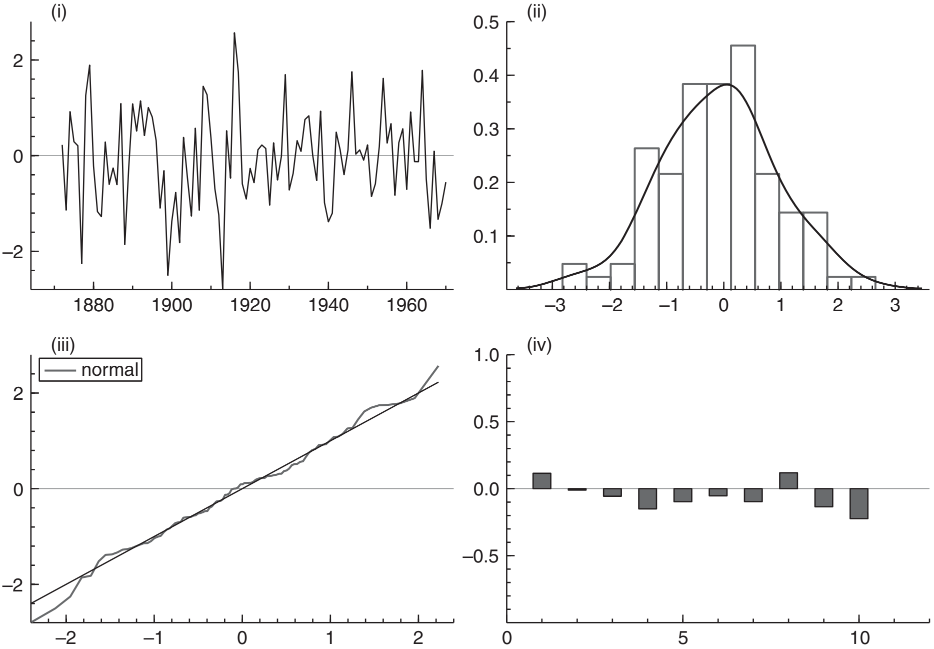

Fig. 2.7 Diagnostic plots for standardised prediction errors: (i) standardised residual; (ii) histogram plus estimated density; (iii) ordered residuals; (iv) correlogram.

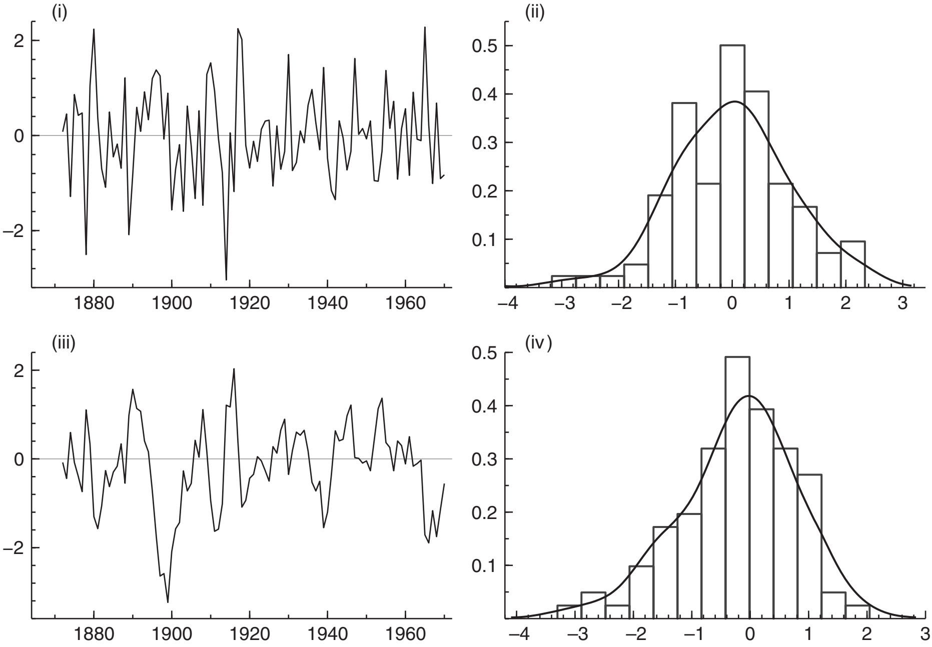

Fig. 2.8 Diagnostic plots for auxiliary residuals: (i) observation residual u_t^*; (ii) histogram and estimated density for u_t^*; (iii) state residual r_t^*; (iv) histogram and estimated density for r_t^*.

The low value for the heteroscedasticity statistic H indicates a degree of heteroscedasticity in the residuals. This is apparent in the plots of u_t^* and r_t^* together with their histograms in Fig. 2.8. These diagnostic plots indicate outliers in 1913 and 1918 and a level break in 1899. The plot of the Nile data confirms these findings.

2.13 Exercises

2.13.1

Consider the local level model (2.3).

(a) Give a model representation for $x_t = y_t - y_{t-1}$, for $t = 2, \ldots, n$.

(b) Show that the model for x_t in (a) can have the same statistical properties as the model given by $x_t = \xi_t + \theta\,\xi_{t-1}$ where $\xi_t \sim \mathrm{N}(0, \sigma_\xi^2)$ are independent disturbances with variance $\sigma_\xi^2 > 0$ and for some value θ.

(c) For what value of θ, in terms of σ_ε^2 and σ_η^2, are the model representations for x_t in (a) and (b) equivalent? Comment.

2.13.2

(a) Using the derivations as in Subsection 2.4.2, develop backwards recursions for the evaluation of $\text{Cov}(\alpha_{t+1}, \alpha_t | Y_n)$ for $t = n, \ldots, 1$.

(b) Using the derivations as in Subsection 2.5.1 and 2.5.2, develop backwards recursions for the evaluation of $\text{Cov}(\varepsilon_t, \eta_t | Y_n)$ for $t = n, \ldots, 1$.

2.13.3

Consider the loglikelihood expression (2.59) and show that the maximum likelihood estimator of a_1 is given by

$$\hat{a}_1 = \frac{1}{n} \sum_{t=1}^{n} u_t^o,$$

where u_t^o is defined as in (2.45) but obtained from the Kalman filter and smoothing recursions with initialisation $a_1 = 0$. Note that we treat the initial state variance P_1 here as a known and finite value.

3 Linear state space models

3.1 Introduction

The general linear Gaussian state space model can be written in a variety of ways; we shall use the form

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{N}(0, H_t), \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t, & \eta_t &\sim \mathrm{N}(0, Q_t), \qquad t = 1, \dots, n, \end{aligned} \tag{3.1}$$

where y_t is a $p \times 1$ vector of observations called the *observation vector* and α_t is an unobserved $m \times 1$ vector called the *state vector*. The idea underlying the model is that the development of the system over time is determined by α_t according to the second equation of (3.1), but because α_t cannot be observed directly we must base the analysis on observations y_t. The first equation of (3.1) is called the *observation equation* and the second is called the *state equation*. The matrices Z_t, T_t, R_t, H_t and Q_t are initially assumed to be known and the error terms ε_t and η_t are assumed to be serially independent and independent of each other at all time points. Matrices Z_t and T_{t-1} can be permitted to depend on $y_1, \dots, y_{t-1}$. The initial state vector α_1 is assumed to be $\mathrm{N}(a_1, P_1)$ independently of $\varepsilon_1, \dots, \varepsilon_n$ and $\eta_1, \dots, \eta_n$, where a_1 and P_1 are first assumed known; we will consider in Chapter 5 how to proceed in the absence of knowledge of a_1 and P_1. In practice, some or all of the matrices Z_t, H_t, T_t, R_t and Q_t will depend on elements of an unknown parameter vector ψ, the estimation of which will be considered in Chapter 7. The same model is used for a classical and a Bayesian analysis. The general linear state space model is the same as (3.1) except that the error densities are written as $\varepsilon_t \sim (0, H_t)$ and $\eta_t \sim (0, Q_t)$, that is, the normality assumption is dropped.

The first equation of (3.1) has the structure of a linear regression model where the coefficient vector α_t varies over time. The second equation represents a first order vector autoregressive model, the Markovian nature of which accounts for many of the elegant properties of the state space model. The local level model (2.3) considered in the last chapter is a simple special case of (3.1). In many applications R_t is the identity. In others, one could define $\eta_t^* = R_t\eta_t$ and $Q_t^* = R_tQ_tR_t'$ and proceed without explicit inclusion of R_t, thus making the model look simpler. However, if R_t is $m \times r$ with $r < m$ and Q_t is nonsingular, there is an obvious advantage in working with nonsingular η_t rather than singular η_t^*. We assume that R_t is a subset of the columns of I_m; in this case R_t is called a *selection matrix* since it selects the rows of the state equation which have nonzero disturbance terms; however, much of the theory remains valid if R_t is a general $m \times r$ matrix.

Model (3.1) provides a powerful tool for the analysis of a wide range of problems. In this chapter we shall give substance to the general theory to be presented in Chapter 4 by describing a number of important applications of the model to problems in time series analysis and in spline smoothing analysis.

3.2 Univariate structural time series models

A *structural time series model* is one in which the trend, seasonal and error terms in the basic model (2.1), plus other relevant components, are modelled explicitly. In this section we shall consider structural models for the case where y_t is univariate; we shall extend this to the case where y_t is multivariate in Section 3.3. A detailed discussion of structural time series models, together with further references, has been given by Harvey (1989).

3.2.1 Trend component

The local level model considered in Chapter 2 is a simple form of a structural time series model. By adding a slope term ν_t, which is generated by a random walk, we obtain the model

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim \mathrm{N}\big(0, \sigma_\xi^2\big), \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim \mathrm{N}\big(0, \sigma_\zeta^2\big). \end{aligned} \tag{3.2}$$

This is called the *local linear trend* model. If $\xi_t = \zeta_t = 0$ then $\nu_{t+1} = \nu_t = \nu$, say, and $\mu_{t+1} = \mu_t + \nu$ so the trend is exactly linear and (3.2) reduces to the deterministic linear trend plus noise model. The form (3.2) with $\sigma_\xi^2 > 0$ and $\sigma_\zeta^2 > 0$ allows the trend level and slope to vary over time.

Applied workers sometimes complain that the series of values of μ_t obtained by fitting this model does not look smooth enough to represent their idea of what a trend should look like. This objection can be met by setting $\sigma_\xi^2 = 0$ at the outset and fitting the model under this restriction. Essentially the same effect can be obtained by using in place of the second and third equation of (3.2) the model $\Delta^2 \mu_{t+1} = \zeta_t$, i.e. $\mu_{t+1} = 2\mu_t - \mu_{t-1} + \zeta_t$ where Δ is the first difference operator defined by $\Delta x_t = x_t - x_{t-1}$. This and its extension $\Delta^r \mu_t = \zeta_t$ for $r > 2$ have been advocated for modelling trend in state space models in a series of papers by Young and his collaborators under the name *integrated random walk* models; see, for example, Young, Lane, Ng and Palmer (1991). We see that (3.2) can be written in the form

$$y_t = (1 \quad 0)\begin{pmatrix}\mu_t \\ \nu_t\end{pmatrix} + \varepsilon_t,$$

$$\begin{pmatrix}\mu_{t+1} \\ \nu_{t+1}\end{pmatrix} = \begin{bmatrix}1 & 1 \\ 0 & 1\end{bmatrix}\begin{pmatrix}\mu_t \\ \nu_t\end{pmatrix} + \begin{pmatrix}\xi_t \\ \zeta_t\end{pmatrix},$$

which is a special case of (3.1).

3.2.2 Seasonal component

To model the seasonal term γ_t in (2.1), suppose there are s 'months' per 'year'. Thus for monthly data $s = 12$, for quarterly data $s = 4$ and for daily data, when modelling the weekly pattern, $s = 7$. If the seasonal pattern is constant over time, the seasonal values for months 1 to s can be modelled by the constants $\gamma_1^*, \dots, \gamma_s^*$ where $\sum_{j=1}^{s} \gamma_j^* = 0$. For the jth 'month' in 'year' i we have $\gamma_t = \gamma_j^*$ where $t = s(i-1)+j$ for $i = 1, 2, \dots$ and $j = 1, \dots, s$. It follows that $\sum_{j=0}^{s-1} \gamma_{t+1-j} = 0$ so $\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j}$ with $t = s-1, s, \dots$. In practice we often wish to allow the seasonal pattern to change over time. A simple way to achieve this is to add an error term ω_t to this relation giving the model

$$\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j} + \omega_t, \qquad \omega_t \sim \mathrm{N}(0, \sigma_\omega^2), \tag{3.3}$$

for $t = 1, \dots, n$ where initialisation at $t = 1, \dots, s-1$ will be taken care of later by our general treatment of the initialisation question in Chapter 5. An alternative suggested by Harrison and Stevens (1976) is to denote the effect of season j at time t by γ_{jt} and then let γ_{jt} be generated by the quasi-random walk

$$\gamma_{j,t+1} = \gamma_{jt} + \omega_{jt}, \qquad t = (i-1)s + j, \qquad i = 1, 2, \dots, \qquad j = 1, \dots, s, \tag{3.4}$$

with an adjustment to ensure that each successive set of s seasonal components sums to zero; see Harvey (1989, §2.3.4) for details of the adjustment.

It is often preferable to express the seasonal in a trigonometric form, one version of which, for a constant seasonal, is

$$\gamma_t = \sum_{j=1}^{[s/2]} (\tilde{\gamma}_j \cos \lambda_j t + \tilde{\gamma}_j^* \sin \lambda_j t), \qquad \lambda_j = \frac{2\pi j}{s}, \qquad j = 1, \dots, [s/2], \tag{3.5}$$

where $[a]$ is the largest integer $\leq a$ and where the quantities $\tilde{\gamma}_j$ and $\tilde{\gamma}_j^*$ are given constants. For a time-varying seasonal this can be made stochastic by replacing $\tilde{\gamma}_j$ and $\tilde{\gamma}_j^*$ by the random walks

$$\tilde{\gamma}_{j,t+1} = \tilde{\gamma}_{jt} + \tilde{\omega}_{jt}, \quad \tilde{\gamma}_{j,t+1}^* = \tilde{\gamma}_{jt}^* + \tilde{\omega}_{jt}^*, \quad j = 1, \dots, [s/2], \quad t = 1, \dots, n, \tag{3.6}$$

where $\tilde{\omega}_{jt}$ and $\tilde{\omega}^*_{jt}$ are independent $\mathrm{N}(0, \sigma^2_\omega)$ variables; for details see Young, Lane, Ng and Palmer (1991). An alternative trigonometric form is the quasi-random walk model

$$\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{jt}, \tag{3.7}$$

where

$$\begin{aligned} \gamma_{j,t+1} &= \gamma_{jt} \cos \lambda_j + \gamma^*_{jt} \sin \lambda_j + \omega_{jt}, \\ \gamma^*_{j,t+1} &= -\gamma_{jt} \sin \lambda_j + \gamma^*_{jt} \cos \lambda_j + \omega^*_{jt}, \qquad j = 1, \dots, [s/2], \end{aligned} \tag{3.8}$$

in which the ω_{jt} and ω^*_{jt} terms are independent $\mathrm{N}(0, \sigma^2_\omega)$ variables. We can show that when the stochastic terms in (3.8) are zero, the values of γ_t defined by (3.7) are periodic with period s by taking

$$\begin{aligned} \gamma_{jt} &= \tilde{\gamma}_j \cos \lambda_j t + \tilde{\gamma}^*_j \sin \lambda_j t, \\ \gamma^*_{jt} &= -\tilde{\gamma}_j \sin \lambda_j t + \tilde{\gamma}^*_j \cos \lambda_j t, \end{aligned}$$

which are easily shown to satisfy the deterministic part of (3.8). The required result follows since γ_t defined by (3.5) is periodic with period s. In effect, the deterministic part of (3.8) provides a recursion for (3.5).

The advantage of (3.7) over (3.6) is that the contributions of the errors ω_{jt} and ω^*_{jt} are not amplified in (3.7) by the trigonometric functions $\cos \lambda_j t$ and $\sin \lambda_j t$. We regard (3.3) as the main time domain model and (3.7) as the main frequency domain model for the seasonal component in structural time series analysis. A more detailed discussion of seasonal models is presented in Proietti (2000). In particular, he shows that the seasonal model in trigonometric form with specific variance restrictions for ω_{jt} and ω^*_{jt}, is equivalent to the quasi-random walk seasonal model (3.4).

3.2.3 Basic structural time series model

Each of the four seasonal models of the previous subsection can be combined with either of the trend models to give a structural time series model and all these can be put in the state space form (3.1). For example, for the local linear trend model (3.2) together with model (3.3) we have the observation equation

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \qquad t = 1, \dots, n. \tag{3.9}$$

To represent the model in state space form, we take the state vector as

$$\alpha_t = (\mu_t \quad \nu_t \quad \gamma_t \quad \gamma_{t-1} \quad \dots \quad \gamma_{t-s+2})',$$

and take the system matrices as

$$\begin{aligned} Z_t &= \left(Z_{[\mu]}, Z_{[\gamma]}\right), & T_t &= \operatorname{diag}\left(T_{[\mu]}, T_{[\gamma]}\right), \\ R_t &= \operatorname{diag}\left(R_{[\mu]}, R_{[\gamma]}\right), & Q_t &= \operatorname{diag}\left(Q_{[\mu]}, Q_{[\gamma]}\right), \end{aligned} \tag{3.10}$$

where

$$Z_{[\mu]} = (1, 0), \qquad Z_{[\gamma]} = (1, 0, \dots, 0),$$

$$T_{[\mu]} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad T_{[\gamma]} = \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 1 & 0 \end{bmatrix},$$

$$R_{[\mu]} = I_2, \qquad R_{[\gamma]} = (1, 0, \dots, 0)',$$

$$Q_{[\mu]} = \begin{bmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}, \qquad Q_{[\gamma]} = \sigma_\omega^2.$$

This model plays a prominent part in the approach of Harvey (1989) to structural time series analysis; he calls it the *basic structural time series model*. The state space form of this basic model with $s = 4$ is therefore

$$\alpha_t = (\mu_t \quad \nu_t \quad \gamma_t \quad \gamma_{t-1} \quad \gamma_{t-2})',$$

$$Z_t = (1 \quad 0 \quad 1 \quad 0 \quad 0), \qquad T_t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$R_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad Q_t = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 \\ 0 & 0 & \sigma_\omega^2 \end{bmatrix}.$$

Alternative seasonal specifications can also be used within the basic structural model. The Harrison and Stevens (1976) seasonal model referred to below (3.3) has the $(s + 2) \times 1$ state vector

$$\alpha_t = (\ \mu_t \quad \nu_t \quad \gamma_t \quad \dots \quad \gamma_{t-s+1}\ )',$$

where the relevant parts of the system matrices for substitution in (3.10) are given by

$$Z_{[\gamma]} = (1, 0, \dots, 0), \qquad T_{[\gamma]} = \begin{bmatrix} 0 & I_{s-1} \\ 1 & 0 \end{bmatrix},$$

$$R_{[\gamma]} = I_s, \qquad Q_{[\gamma]} = \sigma_\omega^2 (I_s - \mathtt{1}\mathtt{1}'/s),$$

in which 1 is an $s \times 1$ vector of ones, $\omega_{1t} + \cdots + \omega_{s,t} = 0$ and variance matrix $Q_{[\gamma]}$ has rank $s - 1$.

The seasonal component in trigonometric form (3.8) can be incorporated in the basic structural model with the $(s+1) \times 1$ state vector

$$\alpha_t = (\mu_t \quad \nu_t \quad \gamma_{1t} \quad \gamma_{1t}^* \quad \gamma_{2t} \quad \ldots)',$$

and the relevant parts of the system matrices given by

$$\begin{aligned} Z_{[\gamma]} &= (1, 0, 1, 0, 1, \ldots, 1, 0, 1), & T_{[\gamma]} &= \mathrm{diag}(C_1, \quad \ldots, \quad C_{s^*}, \quad -1), \\ R_{[\gamma]} &= I_{s-1}, & Q_{[\gamma]} &= \sigma_\omega^2 I_{s-1}. \end{aligned}$$

When we assume that s is even, we have $s^* = s\,/\,2$ and

$$C_j = \begin{bmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{bmatrix}, \qquad \lambda_j = \frac{2\pi j}{s}, \qquad j = 1, \ldots, s^*. \tag{3.11}$$

When s is odd, we have $s^* = (s-1)\,/\,2$ and

$$\begin{aligned} Z_{[\gamma]} &= (1, 0, 1, 0, 1, \ldots, 1, 0), & T_{[\gamma]} &= \mathrm{diag}(\, C_1, \quad \ldots, \quad C_{s^*} \,), \\ R_{[\gamma]} &= I_{s-1}, & Q_{[\gamma]} &= \sigma_\omega^2 I_{s-1}. \end{aligned}$$

where C_j is defined in (3.11) for $j = 1, \ldots, s^*$.

3.2.4 Cycle component

Another important component in some time series is the *cycle* c_t which we can introduce by extending the basic time series model (2.2) to

$$y_t = \mu_t + \gamma_t + c_t + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{3.12}$$

In its simplest form c_t is a pure sine wave generated by the relation

$$c_t = \tilde{c}\cos\lambda_c t + \tilde{c}^*\sin\lambda_c t,$$

where λ_c is the frequency of the cycle; the period is $2\pi/\lambda_c$ which is normally substantially greater that the seasonal period s. As with the seasonal, we can allow the cycle to change stochastically over time by means of the relations analogous to (3.8)

$$\begin{aligned} c_{t+1} &= c_t\cos\lambda_c + c_t^*\sin\lambda_c + \tilde{\omega}_t, \\ c_{t+1}^* &= -c_t\sin\lambda_c + c_t^*\cos\lambda_c + \tilde{\omega}_t^*, \end{aligned}$$

where $\tilde{w}_t$ and $\tilde{w}_t^*$ are independent $\mathrm{N}(0, \sigma_{\tilde{w}}^2)$ variables. Cycles of this form fit naturally into the structural time series model framework. The frequency λ_c can be treated as an unknown parameter to be estimated.

The state space representation for the cycle component is similar to a single trigonometric seasonal component but with frequency λ_c. The relevant system matrices for the cycle component are therefore given by

$$\begin{aligned} Z_{[c]} &= (1, 0), & T_{[c]} &= C_c, \\ R_{[c]} &= I_2, & Q_{[c]} &= \sigma_{\tilde{\omega}}^2 I_2, \end{aligned}$$

where the 2×2 matrix C_c is defined as C_j is in (3.11) but with $\lambda_j = \lambda_c$.

In economic time series the cycle component is usually associated with the business cycle. The definition of a business cycle from Burns and Mitchell (1946, p. 3) is typically adopted: 'A cycle consists of expansions occurring at about the same time in many economic activities, followed by similar general recessions, contractions, and revivals which merge into the expansion phase of the next cycle; this sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years; they are not divisible into shorter cycles of similar character with amplitudes approximating their own.' To let our cycle component resemble this definition, we allow its period $2\pi \,/\, \lambda_c$ to range between 1.5 and 12 years and we specify the cycle as a stationary stochastic process. The system matrix C_c for an economic analysis is then given by

$$C_c = \rho_c \begin{bmatrix} \cos \lambda_c & \sin \lambda_c \\ -\sin \lambda_c & \cos \lambda_c \end{bmatrix}, \qquad 1.5 \leq 2\pi \,/\, \lambda_c \leq 12,$$

with damping factor $0 < \rho_c < 1$. We can now define the economic cycle component as

$$c_t = Z_{[c]} \begin{pmatrix} c_t \\ c_t^* \end{pmatrix}, \quad \begin{pmatrix} c_{t+1} \\ c_{t+1}^* \end{pmatrix} = C_c \begin{pmatrix} c_t \\ c_t^* \end{pmatrix} + \begin{pmatrix} \tilde{\omega}_t \\ \tilde{\omega}_t^* \end{pmatrix}, \quad \begin{pmatrix} \tilde{\omega}_t \\ \tilde{\omega}_t^* \end{pmatrix} \sim \mathrm{N}(0, Q_{[c]}), \tag{3.13}$$

for $t = 1, \ldots, n$. When we fit a model with this cycle component to a macroeconomic time series with values for ρ_c and λ_c within their admissable regions, it is justified to interpret the cycle c_t as a business cycle component.

3.2.5 Explanatory variables and intervention effects

Explanatory variables and intervention effects are easily allowed for in the structural model framework. Suppose we have k regressors $x_{1t}, \ldots, x_{kt}$ with regression coefficients $\beta_1, \ldots, \beta_k$ which are constant over time and that we also wish to measure the change in level due to an intervention at time τ. We define an *intervention variable* w_t as follows:

$$\begin{aligned} w_t &= 0, & t &< \tau, \\ &= 1, & t &\geq \tau. \end{aligned}$$

Adding these to the model (3.12) gives

$$y_t = \mu_t + \gamma_t + c_t + \sum_{j=1}^{k} \beta_j x_{jt} + \delta w_t + \varepsilon_t, \qquad t = 1, \dots, n. \tag{3.14}$$

We see that δ measures the change in the level of the series at a known time τ due to an intervention at time τ. The resulting model can readily be put into state space form. For example, if $\gamma_t = c_t = \delta = 0$, $k = 1$ and if μ_t is determined by a local level model, we can take

$$\begin{aligned} \alpha_t &= (\mu_t \quad \beta_{1t})', & Z_t &= (1 \quad x_{1t}), \\ T_t &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad R_t = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & Q_t &= \sigma_\xi^2, \end{aligned}$$

in (3.1). Here, although we have attached a suffix t to β_1 it is made to satisfy $\beta_{1,t+1} = \beta_{1t}$ so it is constant. Other examples of intervention variables are the *pulse intervention variable* defined by

$$\begin{aligned} w_t &= 0, \qquad t < \tau, \qquad t > \tau, \\ &= 1, \qquad t = \tau, \end{aligned}$$

and the *slope intervention variable* defined by

$$\begin{aligned} w_t &= 0, & t &< \tau, \\ &= 1 + t - \tau, & t &\geq \tau. \end{aligned}$$

For other forms of intervention variable designed to represent a more gradual change of level or a transient change see Box and Tiao (1975). Coefficients such as δ which do not change over time can be incorporated into the state vector by setting the corresponding state errors equal to zero. Regression coefficients β_{jt} which change over time can be handled straightforwardly in the state space framework by modelling them by random walks of the form

$$\beta_{j,t+1} = \beta_{jt} + \chi_{jt}, \qquad \chi_{jt} \sim \mathrm{N}\big(0, \sigma_\chi^2\big), \qquad j = 1, \dots, k. \tag{3.15}$$

An example of the use of model (3.14) for intervention analysis is given by Harvey and Durbin (1986) who used it to measure the effect of the British seat belt law on road traffic casualities. Of course, if the cycle term, the regression term or the intervention term are not required, they can be omitted from (3.14). Instead of including regression and intervention coefficients in the state vector, an alternative way of dealing with them is to concentrate them out of the likelihood function and estimate them via regression, as we will show in Subsection 6.2.3.

3.2.6 STAMP

A wide-ranging discussion of structural time series models can be found in Harvey (1989). Supplementary sources for further applications and later work are Harvey and Shephard (1993), Harvey (2006), and Harvey and Koopman (2009). The computer package *STAMP* 8.3 of Koopman, Harvey, Doornik and Shephard (2010) is designed to analyse, model and forecast time series based on univariate and multivariate structural time series models. The package has implemented the Kalman filter and associated algorithms leaving the user free to concentrate on the important part of formulating a model. *STAMP* is a commercial product and more information on it can be obtained from the Internet at

`http://stamp-software.com/`

3.3 Multivariate structural time series models

The methodology of structural time series models lends itself easily to generalisation to multivariate time series. Consider the local level model for a $p \times 1$ vector of observations y_t, that is

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \\ \mu_{t+1} &= \mu_t + \eta_t, \end{aligned} \tag{3.16}$$

where μ_t, ε_t and η_t are $p \times 1$ vectors and

$$\varepsilon_t \sim \mathrm{N}(0, \Sigma_\varepsilon), \qquad \eta_t \sim \mathrm{N}(0, \Sigma_\eta),$$

with $p \times p$ variance matrices Σ_ε and Σ_η. In this so-called *seemingly unrelated time series equations* model, each series in y_t is modelled as in the univariate case, but the disturbances may be correlated instantaneously across series. In the case of a model with other components such as slope, cycle and seasonal, the disturbances associated with the components become vectors which have $p \times p$ variance matrices. The link across the p different time series is through the correlations of the disturbances driving the components.

3.3.1 Homogeneous models

A seemingly unrelated time series equations model is said to be *homogeneous* when the variance matrices associated with the different disturbances are proportional to each other. For example, the homogeneity restriction for the multivariate local level model is

$$\Sigma_\eta = q\Sigma_\varepsilon,$$

where scalar q is the signal-to-noise ratio. This means that all the series in y_t, and linear combinations thereof, have the same dynamic properties which implies that they have the same autocorrelation function for the stationary form of the model. A homogeneous model is a rather restricted model but it is easy to estimate. For further details we refer to Harvey (1989, Chapter 8).

3.3.2 Common levels

Consider the multivariate local level model without the homogeneity restriction but with the assumption that the rank of Σ_η is $r < p$. The model then contains only r underlying level components. We may refer to these as *common levels*. Recognition of such common factors yields models which may not only have an interesting interpretation, but may also provide more efficient inferences and forecasts. With an appropriate ordering of the series the model may be written as

$$\begin{aligned} y_t &= a + A\mu_t^* + \varepsilon_t, \\ \mu_{t+1}^* &= \mu_t^* + \eta_t^*, \end{aligned}$$

where μ_t^* and η_t^* are $r \times 1$ vectors, a is a $p \times 1$ vector and A is a $p \times r$ matrix. We further assume that

$$a = \begin{pmatrix} 0 \\ a^* \end{pmatrix}, \qquad A = \begin{bmatrix} I_r \\ A^* \end{bmatrix}, \qquad \eta_t^* \sim \mathrm{N}(0, \Sigma_\eta^*),$$

where a^* is a $(p-r) \times 1$ vector and A^* is a $(p-r) \times r$ matrix of nonzero values and where variance matrix Σ_η^* is a $r \times r$ positive definite matrix. The matrix A may be interpreted as a factor loading matrix. When there is more than one common factor ($r > 1$), the factor loadings are not unique. A factor rotation may give components with a more interesting interpretation.

The introduction of common factors can also be extended to other multivariate components such as slope, cycle and seasonal. For example, an illustration of a common business cycle component in a model for a vector of economic time series is given by Valle e Azevedo, Koopman and Rua (2006). Further discussions of multivariate extensions of structural time series models are given by Harvey (1989, Chapter 8) and, Harvey and Koopman (1997) and Koopman, Ooms and Hindrayanto (2009).

3.3.3 Latent risk model

Risk is at the centre of many policy decisions in companies, governments and financial institutions. In risk analysis, measures of exposure to risk, outcomes (or events), and, possibly, losses are analysed. Risk itself cannot be observed. For example, in the case of road safety research, exposure is the number of cars (or the number of kilometres travelled), outcome is the number of accidents and loss is the cost of damages (or the number of fatalities). From these measures we can learn about risk and its associated severity. In a time series context, we typically observe the measures in totals for a country (or a region, or a specific group) and for a specific period (month, quarter or other). The time series for exposure y_{1t}, outcome y_{2t} and loss y_{3t} are typically observed with error and are subject to, possibly, trend, seasonal, cycle and regression effects.

We can carry out a time series analysis with direct interpretable variables via the multiplicative latent risk model

$$y_{1t} = \mu_{1t} \times \xi_{1t}, \qquad y_{2t} = \mu_{1t} \times \mu_{2t} \times \xi_{2t}, \qquad y_{3t} = \mu_{1t} \times \mu_{2t} \times \mu_{3t} \times \xi_{3t},$$

where the unobserved components μ_{1t} is exposure corrected for observation error, μ_{2t} is risk and μ_{3t} is severity while ξ_{jt} are the multiplicative errors with their means equal to one, for $j = 1, 2, 3$. The model implies that expected outcome is exposure times risk while expected loss is outcome times severity. By taking logs we obtain the trivariate latent risk model in additive form

$$\begin{pmatrix} \log y_{1t} \\ \log y_{2t} \\ \log y_{3t} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \theta_t + \varepsilon_t,$$

with the signal vector $\theta_t = (\log \mu_{1t}\,, \log \mu_{2t}\,, \log \mu_{3t})'$ and with the disturbance vector $\varepsilon_t = (\log \xi_{1t}\,, \log \xi_{2t}\,, \log \xi_{3t})'$. The signals for exposure, risk and severity (all three variables in logs) can be modelled simultaneously as linear functions of the state vector. The signal vector for our purpose can be given by

$$\theta_t = S_t \alpha_t, \qquad t = 1, \ldots, n,$$

where S_t is a $3 \times m$ system matrix that relates the signal θ_t with the state vector α_t. The model is placed in state space form (3.1) with

$$Z_t = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} S_t.$$

Multiple measures for exposure, outcome and/or loss can be incorporated in this framework in a straightforward manner. Typical applications of the latent risk model are for studies on insurance claims, credit card purchases and road safety. Bijleveld, Commandeur, Gould and Koopman (2008) provide further discussions on the latent risk model and show how the general methodology can be effectively used in the assessment of risk.

3.4 ARMA models and ARIMA models

Autoregressive integrated moving average (ARIMA) time series models were employed for this purpose by Box and Jenkins in their pathbreaking (1970) book; see Box, Jenkins and Reinsel (1994) for the current version of this book. As with structural time series models considered in Section 3.2, Box and Jenkins typically regarded a univariate time series y_t as made up of trend, seasonal and irregular components. However, instead of modelling the various components separately, their idea was to eliminate the trend and seasonal by differencing at the outset of the analysis. The resulting differenced series are treated as a stationary time series, that is, a series where characteristic properties such as means, covariances and so on remain invariant under translation through time. Let $\Delta y_t = y_t - y_{t-1}$, $\Delta^2 y_t = \Delta(\Delta y_t)$, $\Delta_s y_t = y_t - y_{t-s}$, $\Delta_s^2 y_t = \Delta_s(\Delta_s y_t)$, and so on, where we are assuming that we have s 'months' per 'year'. Box and Jenkins suggest that differencing is continued until trend and seasonal effects have been eliminated, giving a new variable $y_t^* = \Delta^d \Delta_s^D y_t$ for $d, D = 0, 1, \ldots$, which we model as a stationary autoregressive moving average ARMA(p, q) model given by

$$y_t^* = \phi_1 y_{t-1}^* + \cdots + \phi_p y_{t-p}^* + \zeta_t + \theta_1 \zeta_{t-1} + \cdots + \theta_q \zeta_{t-q}, \qquad \zeta_t \sim \mathrm{N}\big(0, \sigma_\zeta^2\big), \tag{3.17}$$

with non-negative integers p and q and where ζ_t is a serially independent series of $\mathrm{N}(0, \sigma_\zeta^2)$ disturbances. This can be written in the form

$$y_t^* = \sum_{j=1}^{r} \phi_j y_{t-j}^* + \zeta_t + \sum_{j=1}^{r-1} \theta_j \zeta_{t-j}, \qquad t = 1, \ldots, n, \tag{3.18}$$

where $r = \max(p, q+1)$ and for which some coefficients are zero. Box and Jenkins normally included a constant term in (3.18) but for simplicity we omit this; the modifications needed to include it are straightforward. We use the symbols d, p and q here and elsewhere in their familiar ARIMA context without prejudice to their use in different contexts in other parts of the book.

We now demonstrate how to put these models into state space form, beginning with the case where $d = D = 0$, that is, no differencing is needed, so we can model the series by (3.18) with y_t^* replaced by y_t. Take

$$Z_t = (1 \quad 0 \quad 0 \quad \cdots \quad 0),$$

$$\alpha_t = \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \cdots + \phi_r y_{t-r+1} + \theta_1 \zeta_t + \cdots + \theta_{r-1} \zeta_{t-r+2} \\ \phi_3 y_{t-1} + \cdots + \phi_r y_{t-r+2} + \theta_2 \zeta_t + \cdots + \theta_{r-1} \zeta_{t-r+3} \\ \vdots \\ \phi_r y_{t-1} + \theta_{r-1} \zeta_t \end{pmatrix}, \tag{3.19}$$

and write the state equation for α_{t+1} as in (3.1) with

$$T_t = T = \begin{bmatrix} \phi_1 & 1 & & 0 \\ \vdots & & \ddots & \\ \phi_{r-1} & 0 & & 1 \\ \phi_r & 0 & \cdots & 0 \end{bmatrix}, \qquad R_t = R = \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{pmatrix}, \qquad \eta_t = \zeta_{t+1}. \tag{3.20}$$

This, together with the observation equation $y_t = Z_t \alpha_t$, is equivalent to (3.18) but is now in the state space form (3.1) with $\varepsilon_t = 0$, implying that $H_t = 0$. For example, with $r = 2$ we have the state equation

$$\begin{pmatrix} y_{t+1} \\ \phi_2 y_t + \theta_1 \zeta_{t+1} \end{pmatrix} = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \theta_1 \zeta_t \end{pmatrix} + \begin{pmatrix} 1 \\ \theta_1 \end{pmatrix} \zeta_{t+1}.$$

The form given is not the only state space version of an ARMA model but is a convenient one.

We now consider the case of a univariate nonseasonal nonstationary ARIMA model of order p, d and q, with $d > 0$, given by (3.17) with $y_t^* = \Delta^d y_t$. As an example, we first consider the state space form of the ARIMA model with $p = 2$, $d = 1$ and $q = 1$ which is given by

$$
\begin{aligned}
y_t &= (1 \quad 1 \quad 0)\alpha_t, \\
\alpha_{t+1} &= \begin{bmatrix} 1 & 1 & 0 \\ 0 & \phi_1 & 1 \\ 0 & \phi_2 & 0 \end{bmatrix} \alpha_t + \begin{pmatrix} 0 \\ 1 \\ \theta_1 \end{pmatrix} \zeta_{t+1},
\end{aligned}
$$

with the state vector defined as

$$
\alpha_t = \begin{pmatrix} y_{t-1} \\ y_t^* \\ \phi_2 y_{t-1}^* + \theta_1 \zeta_t \end{pmatrix},
$$

and $y_t^* = \Delta y_t = y_t - y_{t-1}$. This example generalises easily to ARIMA models with $d = 1$ with other values for p and q. The ARIMA model with $p = 2$, $d = 2$ and $q = 1$ in state space form is given by

$$
\begin{aligned}
y_t &= (1 \quad 1 \quad 1 \quad 0)\alpha_t, \\
\alpha_{t+1} &= \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & \phi_1 & 1 \\ 0 & 0 & \phi_2 & 0 \end{bmatrix} \alpha_t + \begin{pmatrix} 0 \\ 0 \\ 1 \\ \theta_1 \end{pmatrix} \zeta_{t+1},
\end{aligned}
$$

with

$$
\alpha_t = \begin{pmatrix} y_{t-1} \\ \Delta y_{t-1} \\ y_t^* \\ \phi_2 y_{t-1}^* + \theta_1 \zeta_t \end{pmatrix},
$$

and $y_t^* = \Delta^2 y_t = \Delta(y_t - y_{t-1})$. The relations between y_t, Δy_t and $\Delta^2 y_t$ follow immediately since

$$
\begin{aligned}
\Delta y_t &= \Delta^2 y_t + \Delta y_{t-1}, \\
y_t &= \Delta y_t + y_{t-1} = \Delta^2 y_t + \Delta y_{t-1} + y_{t-1}.
\end{aligned}
$$

We deal with the unknown nonstationary values y_0 and Δy_0 in the initial state vector α_1 in Subsection 5.6.3 where we describe the initialisation procedure for filtering and smoothing. Instead of estimating y_0 and Δy_0 directly, we treat these elements of α_1 as diffuse random elements while the other elements, including y_t^*, are stationary which have proper unconditional means and variances. The need to facilitate the initialisation procedure explains why we set up the state space model in this form. The state space forms for ARIMA models with other values for p^*, d and q^* can be represented in similar ways. The advantage of the state space formulation is that the array of techniques that have been developed for state space models are made available for ARMA and ARIMA models. In particular, techniques for exact maximum likelihood estimation and for initialisation are available.

As indicated above, for seasonal series both trend and seasonal are eliminated by the differencing operation $y_t^* = \Delta^d \Delta_s^D y_t$ prior to modelling y_t^* by a stationary ARMA model of the form (3.18). The resulting model for y_t^* can be put into state space form by a straightforward extension of the above treatment. A well-known seasonal ARIMA model is the so-called *airline model* which is given by

$$y_t^* = \Delta\Delta_{12} y_t = \zeta_t - \theta_1 \zeta_{t-1} - \theta_{12}\zeta_{t-12} + \theta_1\theta_{12}\zeta_{t-13}, \tag{3.21}$$

which has a standard ARIMA state space representation.

It is interesting to note that for many state space models an inverse relation holds in the sense that the state space model has an ARIMA representation. For example, if second differences are taken in the local linear trend model (3.2), the terms in μ_t and ν_t disappear and we obtain

$$\Delta^2 y_t = \varepsilon_{t+2} - 2\varepsilon_{t+1} + \varepsilon_t + \xi_{t+1} - \xi_t + \zeta_t.$$

Since the first two autocorrelations of this are nonzero and the rest are zero, we can write it as a moving average series $\zeta_t^* + \theta_1\zeta_{t-1}^* + \theta_2\zeta_{t-2}^*$ where θ_1 and θ_2 are the moving average parameters and the ζ_t^*'s are independent $\mathrm{N}(0, \sigma_{\zeta^*}^2)$ disturbances. In Box and Jenkins' notation this is an ARIMA(0,2,2) model. We obtain the representation

$$\Delta^2 y_t = \zeta_t^* + \theta_1\zeta_{t-1}^* + \theta_2\zeta_{t-2}^*, \tag{3.22}$$

where the local linear trend model imposes a more restricted space on θ_1 and θ_2 than is required for a non-invertible ARIMA(0,2,2) model. A more elaborate discussion on these issues is given by Harvey (1989, §2.5.3).

It is important to recognise that the model (3.22) is less informative than (3.2) since it has lost the information that exists in the form (3.2) about the level μ_t and the slope ν_t. If a seasonal term generated by model (3.3) is added to the local linear trend model, the corresponding ARIMA model has the form

$$\Delta^2\Delta_s y_t = \zeta_t^* + \sum_{j=1}^{s+2} \theta_j \zeta_{t-j}^*,$$

where $\theta_1, \ldots, \theta_{s+2}$ are determined by the four variances σ_ε^2, σ_ξ^2, σ_ζ^2 and σ_ω^2. In this model, information about the seasonal is lost as well as information about the trend. The fact that structural time series models provide explicit information about trend and seasonal, whereas ARIMA models do not, is an important advantage that the structural modelling approach has over ARIMA modelling. We shall make a detailed comparison of the two approaches to time series analysis in Subsection 3.10.1.

3.5 Exponential smoothing

In this section we consider the development of exponential smoothing methods in the 1950s and we examine their relation to simple forms of state space and Box–Jenkins models. These methods have been primarily developed for the purpose of forecasting. The term 'smoothing' is used in a somewhat different context in this section and should not be related to the term 'smoothing' as it is used elsewhere in the book, notably in Chapters 2 and 4.

Let us start with the introduction in the 1950s of the exponentially weighted moving average (EWMA) for one-step ahead forecasting of y_{t+1} given a univariate time series $y_t, y_{t-1}, \ldots$. This has the form

$$\hat{y}_{t+1} = (1-\lambda)\sum_{j=0}^{\infty} \lambda^j y_{t-j}, \qquad 0 < \lambda < 1. \tag{3.23}$$

From (3.23) we deduce immediately the recursion

$$\hat{y}_{t+1} = (1-\lambda)y_t + \lambda\hat{y}_t, \tag{3.24}$$

which is used in place of (3.23) for practical computation. This has a simple structure and requires little storage so it was very convenient for the primitive computers available in the 1950s. As a result, EWMA forecasting became very popular in industry, particularly for sales forecasting of many items simultaneously. We call the operation of calculating forecasts by (3.24) *exponential smoothing*.

Denote the one-step ahead forecast error $y_t - \hat{y}_t$ by u_t and substitute in (3.24) with t replaced by $t-1$; this gives

$$y_t - u_t = (1-\lambda)y_{t-1} + \lambda(y_{t-1} - u_{t-1}),$$

that is,

$$\Delta y_t = u_t - \lambda u_{t-1}. \tag{3.25}$$

Taking u_t to be a series of independent $\mathrm{N}(0, \sigma_u^2)$ variables, we see that we have deduced from the EWMA recursion (3.24) the simple ARIMA model (3.25).

An important contribution was made by Muth (1960) who showed that EWMA forecasts produced by the recursion (3.24) are minimum mean square error forecasts in the sense that they minimise $\mathrm{E}(\hat{y}_{t+1} - y_{t+1})^2$ for observations $y_t, y_{t-1}, \dots$ generated by the local level model (2.3), which for convenience we write in the form

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \\ \mu_{t+1} &= \mu_t + \xi_t, \end{aligned} \tag{3.26}$$

where ε_t and ξ_t are serially independent random variables with zero means and constant variances. Taking first differences of observations y_t generated by (3.26) gives

$$\Delta y_t = y_t - y_{t-1} = \varepsilon_t - \varepsilon_{t-1} + \xi_{t-1}.$$

Since ε_t and ξ_t are serially uncorrelated the autocorrelation coefficient of the first lag for Δy_t is nonzero but all higher autocorrelations are zero. This is the autocorrelation function of a moving average model of order one which, with λ suitably defined, we can write in the form

$$\Delta y_t = u_t - \lambda u_{t-1},$$

which is the same as model (3.25).

We observe the interesting point that these two simple forms of state space and ARIMA models produce the same one-step ahead forecasts and that these can be calculated by the EWMA (3.24) which has proven practical value. We can write this in the form

$$\hat{y}_{t+1} = \hat{y}_t + (1 - \lambda)(y_t - \hat{y}_t),$$

which is the Kalman filter for the simple state space model (3.26).

The EWMA was extended by Holt (1957) and Winters (1960) to series containing trend and seasonal. The extension for trend in the additive case is

$$\hat{y}_{t+1} = m_t + b_t,$$

where m_t and b_t are level and slope terms generated by the EWMA type recursions

$$\begin{aligned} m_t &= (1 - \lambda_1) y_t + \lambda_1 (m_{t-1} + b_{t-1}), \\ b_t &= (1 - \lambda_2)(m_t - m_{t-1}) + \lambda_2 b_{t-1}. \end{aligned}$$

In an interesting extension of the results of Muth (1960), Theil and Wage (1964) showed that the forecasts produced by these Holt–Winters recursions are minimum mean square error forecasts for the state space model

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, \\ \nu_{t+1} &= \nu_t + \zeta_t, \end{aligned} \tag{3.27}$$

which is the local linear trend model (3.2). Taking second differences of y_t generated by (3.27), we obtain

$$\Delta^2 y_t = \zeta_{t-2} + \xi_{t-1} - \xi_{t-2} + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}.$$

This is a stationary series with nonzero autocorrelations at lags 1 and 2 but zero autocorrelations elsewhere. It therefore follows the moving average model

$$\Delta^2 y_t = u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2},$$

which is a simple form of ARIMA model.

Adding the seasonal term $\gamma_{t+1} = -\gamma_t - \cdots - \gamma_{t-s+2} + \omega_t$ from (3.3) to the measurement equation of (3.26) gives the model

$$\begin{aligned} y_t &= \mu_t + \gamma_t + \varepsilon_t, \\ \mu_{t+1} &= \mu_t + \xi_t, \\ \gamma_{t+1} &= -\gamma_t - \cdots - \gamma_{t-s+2} + \omega_t, \end{aligned} \tag{3.28}$$

which is a special case of the structural time series models of Section 3.2. Now take first differences and first seasonal differences of (3.28). We find

$$\Delta\Delta_s y_t = \xi_{t-1} - \xi_{t-s-1} + \omega_{t-1} - 2\omega_{t-2} + \omega_{t-3} + \varepsilon_t - \varepsilon_{t-1} - \varepsilon_{t-s} + \varepsilon_{t-s-1}, \tag{3.29}$$

which is a stationary time series with nonzero autocorrelations at lags 1, 2, $s-1$, s and $s+1$. Consider the airline model (3.21) for general s,

$$\Delta\Delta_s y_t = u_t - \theta_1 u_{t-1} - \theta_s u_{t-s} - \theta_1\theta_s u_{t-s-1},$$

which has been found to fit well many economic time series containing trend and seasonal. It has nonzero autocorrelations at lags 1, $s-1$, s and $s+1$. Now the autocorrelation at lag 2 from model (3.28) arises only from $\mathrm{Var}(\omega_t)$ which in most cases in practice is small. Thus when we add a seasonal component to the models we find again a close correspondence between state space and ARIMA models. A slope component ν_t can be added to (3.28) as in (3.27) without significantly affecting the conclusions.

A pattern is now emerging. Starting with EWMA forecasting, which in appropriate circumstances has been found to work well in practice, we have found that there are two distinct types of models, the state space models and the Box–Jenkins ARIMA models which appear to be very different conceptually but which both give minimum mean square error forecasts from EWMA recursions. The explanation is that when the time series has an underlying structure which is sufficiently simple, then the appropriate state space and ARIMA models are essentially equivalent. It is when we move towards more complex structures that the differences emerge. The above discussion has been based on Durbin (2000b, §3).

3.6 Regression models

The regression model for a univariate series y_t is given by

$$y_t = X_t\beta + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, H_t), \tag{3.30}$$

for $t = 1, \ldots, n$, where X_t is the $1 \times k$ regressor vector with exogenous variables, β is the $k \times 1$ vector of regression coefficients and H_t is the known variance that possibly varies with t. This model can be represented in the state space form (3.1) with $Z_t = X_t$, $T_t = I_k$ and $R_t = Q_t = 0$, so that $\alpha_t = \alpha_1 = \beta$. The generalised least squares estimator of the regression coefficient vector β is given by

$$\hat{\beta} = \left(\sum_{t=1}^{n} X_t' H_t^{-1} X_t \right)^{-1} \sum_{t=1}^{n} X_t' H_t^{-1} y_t.$$

When the Kalman filter is applied to the state space model that represents the regression model (3.30), it effectively computes $\hat{\beta}$ in a recursive manner. In this case the Kalman filter reduces to the recursive least squares method as developed by Plackett (1950).

3.6.1 Regression with time-varying coefficients

Suppose that in the linear regression model (3.30) we wish the coefficient vector β to vary over time. A suitable model for this is to replace β in (3.30) by α_t and to permit each coefficient α_{it} to vary according to a random walk $\alpha_{i,t+1} = \alpha_{it} + \eta_{it}$. This gives a state equation for the vector α_t in the form $\alpha_{t+1} = \alpha_t + \eta_t$. Since the model is a special case of (3.1) with $Z_t = X_t$, $T_t = R_t = I_k$ and Q_t is the known (diagonal) variance matrix for η_t, it can be handled in a routine fashion by Kalman filter and smoothing techniques.

3.6.2 Regression with ARMA errors

Consider a regression model of the form

$$y_t = X_t\beta + \xi_t, \qquad t = 1, \ldots, n, \tag{3.31}$$

where y_t is a univariate dependent variable, X_t is a $1 \times k$ regressor vector, β is its coefficient vector and ξ_t denotes the error which is assumed to follow an ARMA model of form (3.18); this ARMA model may or may not be stationary and some of the coefficients ϕ_j, θ_j may be zero as long as ϕ_r and θ_{r-1} are not both zero. Let α_t be defined as in (3.19) and let

$$\alpha_t^* = \begin{pmatrix} \beta_t \\ \alpha_t \end{pmatrix},$$

where $\beta_t = \beta$. Writing the state equation implied by (3.20) as $\alpha_{t+1} = T\alpha_t + R\eta_t$, let

$$T^* = \begin{bmatrix} I_k & 0 \\ 0 & T \end{bmatrix}, \qquad R^* = \begin{bmatrix} 0 \\ R \end{bmatrix}, \qquad Z_t^* = (X_t \quad 1 \quad 0 \quad \cdots \quad 0),$$

where T and R are defined in (3.20). Then the model

$$y_t = Z_t^* \alpha_t^*, \qquad \alpha_{t+1}^* = T^* \alpha_t^* + R^* \eta_t,$$

is in state space form (3.1) so Kalman filter and smoothing techniques are applicable; these provide an efficient means of fitting model (3.31). It is evident that the treatment can easily be extended to the case where the regression coefficients are determined by random walks as in Subsection 3.6.1. Moreover, with this approach, unlike some others, it is not necessary for the ARMA model used for the errors to be stationary.

3.7 Dynamic factor models

Principal components and factor analysis are widely used statistical methods in the applied social and behavourial sciences. These methods aim to identify commonalities in the covariance structure of high-dimensional data sets. A factor analysis model can be based on the form given by Lawley and Maxwell (1971), that is

$$y_i = \Lambda f_i + u_i, \qquad f_i \sim \mathrm{N}(0, \Sigma_f), \qquad u_i \sim \mathrm{N}(0, \Sigma_u),$$

where y_i represents a zero-mean data vector containing measured characteristics of subject i, for $i = 1, \dots, n$, while Λ is the coefficient matrix for the low-dimensional vector f_i of latent variables. The latent vector f_i with its variance matrix Σ_f and the disturbance vector u_j with its variance matrix Σ_u are assumed to be mutually and serially independent for $i, j = 1, \dots, n$. The factor analysis model implies the decomposition of the variance matrix Σ_y of y_i into

$$\Sigma_y = \Lambda \Sigma_f \Lambda' + \Sigma_u.$$

Estimation of Λ, Σ_f and Σ_u can be carried out by maximum likelihood procedures; the loading coefficients in Λ and the variance matrices are subject to a set of linear restrictions necessary for identification. A detailed discussion of the maximum likelihood approach to factor analysis is given by Lawley and Maxwell (1971).

In the application of factor analysis in a time series context, the measurements in y_i correspond to a time period t rather than a subject i, we therefore have y_t instead of y_i. The time dependence of the measurements can be accounted for by replacing the serially independence assumption for f_t by a serial dependence assumption. For example, we can assume that f_t is modelled by a vector autoregressive process. We can also let f_t depend on the state vector α_t in a linear way, that is, $f_t = U_t\alpha_t$ where U_t is typically a known selection matrix.

In applications to economics, y_t may consist of a large set of macroeconomic indicators associated with variables such as income, consumption, investment and unemployment. These variables are all subject to economic activity that can often be related to the business cycle. The dynamic features of the business cycle may be disentangled into a set of factors with possibly different dynamic characteristics. In the context of finance, y_t may consist of a large set of daily prices or returns from individual stocks that make up the indices such as Standard & Poor's 500 and Dow Jones. A set of underlying factors may represent returns from particular porfolio strategies. In the context of marketing, y_t may contain market shares of groups or sub-groups of product brands. The factors may indicate whether marketing strategies in certain periods affect all market shares or only a selection of market shares. In all these cases it is not realistic to assume that the factors are independent over time so we are required to formulate a dynamic process for the factors. The wide variety of applications of factor analysis in a time series context have led to many contributions in the statistics and econometrics literature on the inference of dynamic factor models; see, for example, Geweke (1977), Engle and Watson (1981), Watson and Engle (1983), Litterman and Scheinkman (1991), Quah and Sargent (1993), Stock and Watson (2002), Diebold, Rudebusch and Aruoba (2006) and Doz, Giannone and Reichlin (2012).

The dynamic factor model can be regarded as a state space model in which the state vector consists of latent factors where dynamic properties are formulated in the state equation of the state space model (3.1). The size of the observation vector is typically large while the dimension of the state vector is small so we have $p >> m$. In the example

$$y_t = \Lambda f_t + \varepsilon_t, \qquad f_t = U_t\alpha_t, \qquad \varepsilon_t \sim \mathrm{N}(0, H_t), \tag{3.32}$$

the dynamic factor model is a special case of the state space model (3.1) for which the observation equation has $Z_t = \Lambda U_t$. In Chapter 6 we will discuss modifications of the general statistical treatment of the state space model in cases where $p >> m$. These modifications are specifically relevant for the inference of dynamic factor analysis and they ensure feasible methods for a state space analysis.

3.8 State space models in continuous time

In contrast to all the models that we have considered so far, suppose that the observation $y(t)$ is a continuous function of time for t in an interval which we take to be $0 \leq t \leq T$. We shall aim at constructing state space models for $y(t)$ which are the analogues in continuous time for models that we have already studied in discrete time. Such models are useful not only for studying phenomena which genuinely operate in continuous time, but also for providing a convenient theoretical base for situations where the observations take place at time points $t_1 \leq \cdots \leq t_n$ which are not equally spaced.

3.8.1 Local level model

We begin by considering a continuous version of the local level model (2.3). To construct this, we need a continuous analogue of the Gaussian random walk. This can be obtained from the *Brownian motion process*, defined as the continuous stochastic process $w(t)$ such that $w(0) = 0$, $w(t) \sim \mathrm{N}(0, t)$ for $0 < t < \infty$, where increments $w(t_2) - w(t_1)$, $w(t_4) - w(t_3)$ for $0 \leq t_1 \leq t_2 \leq t_3 \leq t_4$ are independent. We sometimes need to consider increments $dw(t)$, where $dw(t) \sim \mathrm{N}(0, dt)$ for dt infinitesimally small. Analogously to the random walk $\alpha_{t+1} = \alpha_t + \eta_t$, $\eta_t \sim \mathrm{N}(0, \sigma_\eta^2)$ for the discrete model, we define $\alpha(t)$ by the continuous time relation $d\alpha(t) = \sigma_\eta dw(t)$ where σ_η is an appropriate positive scale parameter. This suggests that as the continuous analogue of the local level model we adopt the continuous time state space model

$$\begin{aligned} y(t) &= \alpha(t) + \varepsilon(t), \\ \alpha(t) &= \alpha(0) + \sigma_\eta w(t), \qquad 0 \leq t \leq T, \end{aligned} \tag{3.33}$$

where $T > 0$.

The nature of $\varepsilon(t)$ in (3.33) requires careful thought. It must first be recognised that for any analysis that is performed digitally, which is all that we consider in this book, $y(t)$ cannot be admitted into the calculations as a continuous record; we can only deal with it as a series of values observed at a discrete set of time points $0 \leq t_1 < t_2 < \cdots < t_n \leq T$. Second, $\mathrm{Var}[\varepsilon(t)]$ must be bounded significantly away from zero; there is no point in carrying out an analysis when $y(t)$ is indistinguishably close to $\alpha(t)$. Third, in order to obtain a continuous analogue of the local level model we need to assume that $\mathrm{Cov}[\varepsilon(t_i), \varepsilon(t_j)] = 0$ for observational points t_i, t_j $(i \neq j)$. It is obvious that if the observational points are close together it may be advisable to set up an autocorrelated model for $\varepsilon(t)$, for example a low-order autoregressive model; however, the coefficients of this would have to be put into the state vector and the resulting model would not be a continuous local level model. In order to allow $\mathrm{Var}[\varepsilon(t)]$ to vary over time we assume that $\mathrm{Var}[\varepsilon(t)] = \sigma^2(t)$ where $\sigma^2(t)$ is a non-stochastic function of t that may depend on unknown parameters. We conclude that in place of (3.33) a more appropriate form of the model is

$$\begin{aligned} y(t) &= \alpha(t) + \varepsilon(t), \qquad && t = t_1, \ldots, t_n, \qquad \varepsilon(t_i) \sim \mathrm{N}[0, \sigma^2(t_i)], \\ \alpha(t) &= \alpha(0) + \sigma_\eta w(t), \qquad && 0 \leq t \leq T. \end{aligned} \tag{3.34}$$

We next consider the estimation of unknown parameters by maximum likelihood. Since by definition the likelihood is equal to

$$p[y(t_1)]p[y(t_2)|y(t_1)]\cdots p[y(t_n)|y(t_1),\ldots,y(t_{n-1})],$$

it depends on $\alpha(t)$ only at values $t_1,\ldots,t_n$. Thus for estimation of parameters we can employ the reduced model

$$\begin{aligned} y_i &= \alpha_i + \varepsilon_i, \\ \alpha_{i+1} &= \alpha_i + \eta_i, \qquad i = 1,\ldots,n, \end{aligned} \tag{3.35}$$

where $y_i = y(t_i)$, $\alpha_i = \alpha(t_i)$, $\varepsilon_i = \varepsilon(t_i)$, $\eta_i = \sigma_\eta[w(t_{i+1}) - w(t_i)]$ and where the ε_i's are assumed to be independent. This is a discrete local level model which differs from (2.3) only because the variances of the ε_i's can be unequal; consequently we can calculate the loglikelihood by a slight modification of the method of Subsection 2.10.1 which allows for the variance inequality.

Having estimated the model parameters, suppose that we wish to estimate $\alpha(t)$ at values $t = t_{j_*}$ between t_j and t_{j+1} for $1 \le j < n$. We adjust and extend equations (3.35) to give

$$\begin{aligned} \alpha_{j_*} &= \alpha_j + \eta_j^*, \\ y_{j_*} &= \alpha_{j_*} + \varepsilon_{j_*}, \\ \alpha_{j+1} &= \alpha_{j_*} + \eta_{j_*}^*, \end{aligned} \tag{3.36}$$

where $y_{j_*} = y(t_{j_*})$ is treated as missing, $\eta_j^* = \sigma_\eta[w(t_{j_*}) - w(t_j)]$ and $\eta_{j_*}^* = \sigma_\eta[w(t_{j+1}) - w(t_{j_*})]$. We can now calculate $\mathrm{E}[\alpha_{j_*}|y(t_1),\ldots,y(t_n)]$ and $\mathrm{Var}[\alpha_{j_*}|y(t_1),\ldots,y(t_n)]$ by routine applications of the Kalman filter and smoother for series with missing observations, as described in Section 2.7, with a slight modification to allow for unequal observational error variances.

3.8.2 Local linear trend model

Now let us consider the continuous analogue of the local linear trend model (3.2) for the case where $\sigma_\xi^2 = 0$, so that, in effect, the trend term μ_t is modelled by the relation $\Delta^2\mu_{t+1} = \zeta_t$. For the continuous case, denote the trend by $\mu(t)$ and the slope by $\nu(t)$ by analogy with (3.2). The natural model for the slope is then $d\nu(t) = \sigma_\zeta dw(t)$, where $w(t)$ is standard Brownian motion and $\sigma_\zeta > 0$, which gives

$$\nu(t) = \nu(0) + \sigma_\zeta w(t), \qquad 0 \le t \le T. \tag{3.37}$$

By analogy with (3.2) with $\sigma_\xi^2 = 0$, the model for the trend level is $d\mu(t) = \nu(t)dt$, giving

$$\begin{aligned}\mu(t) &= \mu(0) + \int_0^t \nu(s)\ ds \\ &= \mu(0) + \nu(0)t + \sigma_\zeta \int_0^t w(s)\ ds. \qquad (3.38)\end{aligned}$$

As before, suppose that $y(t)$ is observed at times $t_1 \leq \cdots \leq t_n$. Analogously to (3.34), the observation equation for the continuous model is

$$y(t) = \alpha(t) + \varepsilon(t), \qquad t = t_1, \ldots, t_n, \qquad \varepsilon(t_i) \sim \mathrm{N}[0, \sigma^2(t_i)], \qquad (3.39)$$

and the state equation can be written in the form

$$d\begin{bmatrix} \mu(t) \\ \nu(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \mu(t) \\ \nu(t) \end{bmatrix} dt + \sigma_\zeta \begin{bmatrix} 0 \\ dw(t) \end{bmatrix}. \qquad (3.40)$$

For maximum likelihood estimation we employ the discrete state space model,

$$y_i = \mu_i + \varepsilon_i,$$

$$\begin{pmatrix} \mu_{i+1} \\ \nu_{i+1} \end{pmatrix} = \begin{bmatrix} 1 & \delta_i \\ 0 & 1 \end{bmatrix}\begin{pmatrix} \mu_i \\ \nu_i \end{pmatrix} + \begin{pmatrix} \xi_i \\ \zeta_i \end{pmatrix}, \qquad i = 1, \ldots, n, \qquad (3.41)$$

where $\mu_i = \mu(t_i)$, $\nu_i = \nu(t_i)$, $\varepsilon_i = \varepsilon(t_i)$ and $\delta_i = t_{i+1} - t_i$; also

$$\xi_i = \sigma_\zeta \int_{t_i}^{t_{i+1}} [w(s) - w(t_i)]\, ds,$$

and

$$\zeta_i = \sigma_\zeta[w(t_{i+1}) - w(t_i)],$$

as can be verified from (3.37) and (3.38). From (3.39), $\mathrm{Var}(\varepsilon_i) = \sigma^2(t_i)$. Since $\mathrm{E}[w(s) - w(t_i)] = 0$ for $t_i \leq s \leq t_{i+1}$, $\mathrm{E}(\xi_i) = \mathrm{E}(\zeta_i) = 0$. To calculate $\mathrm{Var}(\xi_i)$, approximate ξ_i by the sum

$$\frac{\delta_i}{M}\sum_{j=0}^{M-1}(M-j)w_j$$

where $w_j \sim \mathrm{N}(0, \sigma_\zeta^2\delta_i/M)$ and $\mathrm{E}(w_j w_k) = 0$ $(j \neq k)$. This has variance

$$\sigma_\zeta^2\frac{\delta_i^3}{M}\sum_{j=0}^{M-1}\left(1 - \frac{j}{M}\right)^2,$$

which converges to

$$\sigma_\zeta^2 \delta_i^3 \int_0^1 x^2 \ dx = \frac{1}{3}\sigma_\zeta^2 \delta_i^3$$

as $M \to \infty$. Also,

$$\begin{aligned} \mathrm{E}(\xi_i \zeta_i) &= \sigma_\zeta^2 \int_{t_i}^{t_{i+1}} \mathrm{E}[\{w(s) - w(t_i)\}\{w(t_{i+1}) - w(t_i)\}]\, ds \\ &= \sigma_\zeta^2 \int_0^{\delta_i} x\, dx \\ &= \frac{1}{2}\sigma_\zeta^2 \delta_i^2, \end{aligned}$$

and $\mathrm{E}(\zeta_i^2) = \sigma_\zeta^2 \delta_i$. Thus the variance matrix of the disturbance term in the state equation (3.41) is

$$Q_i = \mathrm{Var}\begin{pmatrix} \xi_i \\ \zeta_i \end{pmatrix} = \sigma_\zeta^2 \delta_i \begin{bmatrix} \frac{1}{3}\delta_i^2 & \frac{1}{2}\delta_i \\ \frac{1}{2}\delta_i & 1 \end{bmatrix}. \tag{3.42}$$

The loglikelihood is then calculated by means of the Kalman filter as in Section 7.2.

As with model (3.34), adjustments (3.36) can also be introduced to model (3.39) and (3.40) in order to estimate the conditional mean and variance matrix of the state vector $[\mu(t), \nu(t)]'$ at values of t other than $t_1, \ldots, t_n$. Chapter 9 of Harvey (1989) may be consulted for extensions to more general models.

3.9 Spline smoothing

3.9.1 Spline smoothing in discrete time

Suppose we have a univariate series $y_1, \ldots, y_n$ of values which are equispaced in time and we wish to approximate the series by a relatively smooth function $\mu(t)$. A standard approach is to choose $\mu(t)$ by minimising

$$\sum_{t=1}^{n} [y_t - \mu(t)]^2 + \lambda \sum_{t=1}^{n} [\Delta^2 \mu(t)]^2 \tag{3.43}$$

with respect to $\mu(t)$ for given $\lambda > 0$. It is important to note that we are considering $\mu(t)$ here to be a discrete function of t at time points $t = 1, \ldots, n$, in contrast to the situation considered in the next section where $\mu(t)$ is a continuous function of time. If λ is small, the values of $\mu(t)$ will be close to the y_t's but $\mu(t)$ may not be smooth enough. If λ is large the $\mu(t)$ series will be smooth but the values of $\mu(t)$ may not be close enough to the y_t's. The function $\mu(t)$ is called a *spline*. Reviews of methods related to this idea are given in Silverman

(1985), Wahba (1990) and Green and Silverman (1994). Note that in this book we usually take t as the time index but it can also refer to other sequentially ordered measures such as temperature, earnings and speed.

Let us now consider this problem from a state space standpoint. Let $\alpha_t = \mu(t)$ for $t = 1, \ldots, n$ and assume that y_t and α_t obey the state space model

$$y_t = \alpha_t + \varepsilon_t, \qquad \Delta^2 \alpha_t = \zeta_t, \qquad t = 1, \ldots, n, \tag{3.44}$$

where $\mathrm{Var}(\varepsilon_t) = \sigma^2$ and $\mathrm{Var}(\zeta_t) = \sigma^2/\lambda$ with $\lambda > 0$. We observe that the second equation of (3.44) is one of the smooth models for trend considered in Section 3.2. For simplicity suppose that α_{-1} and α_0 are fixed and known. The log of the joint density of $\alpha_1, \ldots, \alpha_n, y_1, \ldots, y_n$ is then, apart from irrelevant constants,

$$-\frac{\lambda}{2\sigma^2} \sum_{t=1}^{n} \left(\Delta_t^2 \alpha_t\right)^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (y_t - \alpha_t)^2. \tag{3.45}$$

Now suppose that our objective is to smooth the y_t series by estimating α_t by $\hat{\alpha}_t = E(\alpha_t | Y_n)$. We shall employ a technique that we shall use extensively later so we state it in general terms. Suppose $\alpha = (\alpha_1', \ldots, \alpha_n')'$ and $y = (y_1', \ldots, y_n')'$ are jointly normally distributed stacked vectors with density $p(\alpha, y)$ and we wish to calculate $\hat{\alpha} = E(\alpha|y)$. Then $\hat{\alpha}$ is the solution of the equations

$$\frac{\partial \log p(\alpha, y)}{\partial \alpha} = 0.$$

This follows since $\log p(\alpha|y) = \log p(\alpha, y) - \log p(y)$ so $\partial \log p(\alpha|y)/\partial\alpha = \partial \log p(\alpha, y)/\partial\alpha$. Now the solution of the equations $\partial \log p(\alpha|y)/\partial\alpha = 0$ is the mode of the density $p(\alpha|y)$ and since the density is normal the mode is equal to the mean vector $\hat{\alpha}$. The conclusion follows. Since $p(\alpha|y)$ is the conditional distribution of α given y, we call this technique *conditional mode estimation* of $\alpha_1, \ldots, \alpha_n$.

Applying this technique to (3.45), we see that $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ can be obtained by minimising

$$\sum_{t=1}^{n} (y_t - \alpha_t)^2 + \lambda \sum_{t=1}^{n} (\Delta^2 \alpha_t)^2.$$

Comparing this with (3.43), and ignoring for the moment the initialisation question, we see that the spline smoothing problem can be solved by finding $E(\alpha_t|Y_n)$ for model (3.44). This is achieved by a standard extension of the smoothing technique of Section 2.4 that will be given in Section 4.4. It follows that state space techniques can be used for spline smoothing. Treatments along these lines have been given by Kohn, Ansley and Wong (1992). This approach has the advantages that the models can be extended to include extra features such as explanatory variables, calendar variations and intervention effects in the ways indicated earlier in this chapter; moreover, unknown quantities, for example λ in (3.43), can be estimated by maximum likelihood using methods that we shall describe in Chapter 7.

3.9.2 Spline smoothing in continuous time

Let us now consider the smoothing problem where the observation $y(t)$ is a continuous function of time t for t in an interval which for simplicity we take to be $0 \leq t \leq T$. Suppose that we wish to smooth $y(t)$ by a function $\mu(t)$ given a sample of values $y(t_i)$ for $i = 1, \dots, n$ where $0 < t_1 < \cdots < t_n < T$. A traditional approach to the problem is to choose $\mu(t)$ to be the twice-differentiable function on $(0, T)$ which minimises

$$\sum_{i=1}^{n} [y(t_i) - \mu(t_i)]^2 + \lambda \int_0^T \left[\frac{\partial^2 \mu(t)}{\partial t^2} \right]^2 dt, \tag{3.46}$$

for given $\lambda > 0$. We observe that (3.46) is the analogue in continuous time of (3.43) in discrete time. This is a well-known problem, a standard treatment to which is presented in Chapter 2 of Green and Silverman (1994). Their approach is to show that the resulting $\mu(t)$ must be a *cubic spline*, which is defined as a cubic polynomial function in t between each pair of time points t_i, t_{i+1} for $i = 0, 1, \dots, n$ with $t_0 = 0$ and $t_{n+1} = T$, such that $\mu(t)$ and its first two derivatives are continuous at each t_i for $i = 1, \dots, n$. The properties of the cubic spline are then used to solve the minimisation problem. In contrast, we shall present a solution based on a continuous time state space model of the kind considered in Section 3.8.

We begin by adopting a model for $\mu(t)$ in the form (3.38), which for convenience we reproduce here as

$$\mu(t) = \mu(0) + \nu(0)t + \sigma_\zeta \int_0^t w(s)\, ds, \qquad 0 \leq t \leq T. \tag{3.47}$$

This is a natural model to consider since it is the simplest model in continuous time for a trend with smoothly varying slope. As the observation equation we take

$$y(t_i) = \mu(t_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \qquad i = 1, \dots, n,$$

where the ε_i's are independent of each other and of $w(t)$ for $0 < t \leq T$. We have taken $\mathrm{Var}(\varepsilon_i)$ to be constant since this is a reasonable assumption for many smoothing problems, and also since it leads to the same solution to the problem of minimising (3.46) as the Green–Silverman approach.

Since $\mu(0)$ and $\nu(0)$ are normally unknown, we represent them by diffuse priors. On these assumptions, Wahba (1978) has shown that on taking

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\zeta^2},$$

the conditional mean $\hat{\mu}(t)$ of $\mu(t)$ defined by (3.47), given the observations $y(t_1), \dots, y(t_n)$, is the solution to the problem of minimising (3.46) with respect to $\mu(t)$. We shall not give details of the proof here but will instead refer to discussions of the result by Wecker and Ansley (1983) and Green and Silverman (1994).

The result is important since it enables problems in spline smoothing to be solved by state space methods. We note that Wahba and Wecker and Ansley in the papers cited consider the more general problem in which the second term of (3.46) is replaced by the more general form

$$\lambda \int_0^T \left[\frac{d^m \mu(t)}{dt^m} \right] dt,$$

for $m = 2, 3, \dots$.

We have reduced the problem of minimising (3.46) to the treatment of a special case of the state space model (3.39) and (3.40) in which $\sigma^2(t_i) = \sigma_\varepsilon^2$ for all i. We can therefore compute $\hat{\mu}(t)$ and $\mathrm{Var}[\mu(t)|y(t_1), \dots, y(t_n)]$ by routine Kalman filtering and smoothing. We can also compute the loglikelihood and, consequently, estimate λ by maximum likelihood; this can be done efficiently by concentrating out σ_ε^2 by a straightforward extension of the method described in Subsection 2.10.2 and then maximising the concentrated loglikelihood with respect to λ in a one-dimensional search. The implication of these results is that the flexibility and computational power of state space methods can be employed to solve problems in spline smoothing.

3.10 Further comments on state space analysis

In this section we provide discussions and illustrations of state space time series analysis. First we compare the state space and Box–Jenkins approaches to time series analysis. Next we provide examples of how problems in time series analysis can be handled within a state space framework.

3.10.1 State space versus Box–Jenkins approaches

The early development of state space methodology took place in the field of engineering rather than statistics, starting with the pathbreaking paper of Kalman (1960). In this paper Kalman did two crucially important things. He showed that a very wide class of problems could be encapsulated in a simple linear model, essentially the state space model (3.1). Secondly he showed how, due to the Markovian nature of the model, the calculations needed for practical application of the model could be set up in recursive form in a way that was particularly convenient on a computer. A huge amount of work was done in the development of these ideas in the engineering field. In the 1960s to the early 1980s contributions to state space methodology from statisticians and econometricians were isolated and sporadic. In recent years however there has been a rapid growth of interest in the field in both statistics and econometrics as is indicated by references throughout the book.

The key advantage of the state space approach is that it is based on a structural analysis of the problem. The different components that make up the series, such as trend, seasonal, cycle and calendar variations, together with the effects of explanatory variables and interventions, are modelled separately before being put together in the state space model. It is up to the investigator to identify and model any features in particular situations that require special treatment. In contrast, the Box–Jenkins approach is a kind of 'black box', in which the model adopted depends purely on the data without prior analysis of the structure of the system that generated the data. A second advantage of state space models is that they are flexible. Because of the recursive nature of the models and of the computational techniques used to analyse them, it is straightforward to allow for known changes in the structure of the system over time. Other advantages of a state space analysis are (i) its treatment of missing observations; (ii) explanatory variables can be incorporated into the model without difficulty; (iii) associated regression coefficients can be permitted to vary stochastically over time if this seems to be called for in the application; (iv) trading-day adjustments and other calendar variations can be readily taken care of; (v) no extra theory is required for forecasting since all that is needed is to project the Kalman filter forward into the future.

When employing the Box–Jenkins approach, the elimination of trend and seasonal by differencing may not be a drawback if forecasting is the only object of the analysis. However, in many contexts, particularly in official statistics and some econometric applications, knowledge about such components has intrinsic importance. It is true that estimates of trend and seasonal can be 'recovered' from the differenced series by maximising the residual mean square as in Burman (1980) but this seems an artificial procedure which is not as appealing as modelling the components directly. Furthermore, the requirement that the differenced series should be stationary is a weakness of the theory. In the economic and social fields, real series are never stationary however much differencing is done. The investigator has to face the question, how close to stationarity is close enough? This is a hard question to answer.

In practice it is found that the airline model and similar ARIMA models fit many data sets quite well, but it can be argued that the reason for this is that they are approximately equivalent to plausible state space models. This point is discussed at length by Harvey (1989, pp. 72–73). As we move away from airline-type models, the model identification process in the Box–Jenkins system becomes difficult to apply. The main tool is the sample autocorrelation function which is notoriously imprecise due to its high sampling variability. Practitioners in applied time series analysis are familiar with the fact that many examples can be found where the data appear to be explained equally well by models whose specifications look very different. The above discussion has been based on Durbin (2000b, §3).

3.10.2 Benchmarking

A common problem in official statistics is the adjustment of monthly or quarterly observations, obtained from surveys and therefore subject to survey errors, to agree with annual totals obtained from censuses and assumed to be free from error. The annual totals are called *benchmarks* and the process is called *benchmarking*. We shall show how the problem can be handled within a state space framework.

Denote the survey observations, which we take to be monthly ($s = 12$), by y_t and the true values they are intended to estimate by y_t^* for $t = 12(i-1) + j$, $i = 1, \ldots, \ell$ and $j = 1, \ldots, 12$, where ℓ is the number of years. Thus the survey error is $y_t - y_t^*$ which we denote by $\sigma_t^s \xi_t^s$ where σ_t^s is the standard deviation of the survey error at time t. The error ξ_t^s is modelled as an AR(1) model with unit variance. In principle, ARMA models of higher order could be used. We assume that the values of σ_t^s are available from survey experts and that the errors are bias free; we will mention the estimation of bias later. The benchmark values are given by $x_i = \sum_{j=1}^{12} y_{12(i-1)+j}^*$ for $i = 1, \ldots, \ell$. We suppose for simplicity of exposition that we have these annual values for all years in the study though in practice the census values will usually lag a year or two behind the survey observations. We take as the model for the observations

$$y_t = \mu_t + \gamma_t + \sum_{j=1}^{k} \delta_{jt} w_{jt} + \varepsilon_t + \sigma_t^s \xi_t^s, \qquad t = 1, \ldots, 12\ell, \tag{3.48}$$

where μ_t is trend, γ_t is seasonal and the term $\sum_{j=1}^{k} \delta_{jt} w_{jt}$ represents systematic effects such as the influence of calendar variations which can have a substantial effect on quantities such as retail sales but which can vary slowly over time.

The series is arranged in the form

$$y_1, \ldots, y_{12}, x_1, y_{13}, \ldots, y_{24}, x_2, y_{25}, \ldots, y_{12\ell}, x_\ell.$$

Let us regard the time point in the series at which the benchmark occurs as $t = (12i)'$; thus the point $t = (12i)'$ occurs in the series between $t = 12i$ and $t = 12i + 1$. It seems reasonable to update the regression coefficients δ_{jt} only once a year, say in January, so we take for these coefficients the model

$$\begin{aligned} \delta_{j,12i+1} &= \delta_{j,12i} + \zeta_{j,12i}, && j = 1, \ldots, k, \qquad i = 1, \ldots, \ell, \\ \delta_{j,t+1} &= \delta_{j,t}, && \text{otherwise.} \end{aligned}$$

Take the integrated random walk model for the trend component and model (3.3) for the seasonal component, that is,

$$\Delta^2 \mu_t = \xi_t, \qquad \gamma_t = -\sum_{j=1}^{11} \gamma_{t-j} + \omega_t;$$

see Section 3.2 for alternative trend and seasonal models. It turns out to be convenient to put the observation errors into the state vector, so we take

$$\alpha_t = \big(\mu_t, \dots, \mu_{t-11}, \gamma_t, \dots, \gamma_{t-11}, \delta_{1t}, \dots, \delta_{kt}, \varepsilon_t, \dots, \varepsilon_{t-11}, \xi_t^s\big)'.$$

Thus $y_t = Z_t \alpha_t$ where

$$Z_t = \big(1, 0, \dots, 0, 1, 0, \dots, 0, w_{1t}, \dots, w_{kt}, 1, 0, \dots, 0, \sigma_t^s\big), \qquad t = 1, \dots, n,$$

and $x_i = Z_t \alpha_t$ where

$$Z_t = \left(1, \dots, 1, 0, \dots, 0, \sum_{s=12i-11}^{12i} w_{1s}, \dots, \sum_{s=12i-11}^{12i} w_{ks}, 1, \dots, 1, 0\right), \; t = (12i)',$$

for $i = 1, \dots, \ell$. Using results from Section 3.2 it is easy to write down the state transition from α_t to α_{t+1} for $t = 12i - 11$ to $t = 12i - 1$, taking account of the fact that $\delta_{j,t+1} = \delta_{jt}$. From $t = 12i$ to $t = (12i)'$ the transition is the identity. From $t = (12i)'$ to $t = 12i + 1$, the transition is the same as for $t = 12i - 11$ to $t = 12i - 1$, except that we take account of the relation $\delta_{j,12i+1} = \delta_{j,12i} + \zeta_{j,12i}$ where $\zeta_{j,12i} \neq 0$.

There are many variants of the benchmarking problem. For example, the annual totals may be subject to error, the benchmarks may be values at a particular month, say December, instead of annual totals, the survey observations may be biased and the bias needs to be estimated, more complicated models than the AR(1) model can be used to model y_t^*; finally, the observations may behave multiplicatively whereas the benchmark constraint is additive, thus leading to a nonlinear model. All these variants are dealt with in a comprehensive treatment of the benchmarking problem by Durbin and Quenneville (1997). They also consider a two-step approach to the problem in which a state space model is first fitted to the survey observations and the adjustments to satisfy the benchmark constraints take place in a second stage.

Essentially, this example demonstrates that the state space approach can be used to deal with situations in which the data come from two different sources. Another example of such problems will be given in Subsection 3.10.3 where we model different series which aim at measuring the same phenomenon simultaneously, which are all subject to sampling error and which are observed at different time intervals.

3.10.3 Simultaneous modelling of series from different sources

A different problem in which data come from two different sources has been considered by Harvey and Chung (2000). Here the objective is to estimate the level of UK unemployment and the month-to-month change of unemployment given two different series. Of these, the first is a series y_t of observations obtained from a monthly survey designed to estimate unemployment according to an internationally accepted standard definition (the so-called ILO definition where ILO stands for International Labour Office); this estimate is subject to survey error. The second series consists of monthly counts x_t of the number of individuals claiming unemployment benefit; although these counts are known accurately, they do not themselves provide an estimate of unemployment consistent with the ILO definition. Even though the two series are closely related, the relationship is not even approximately exact and it varies over time. The problem to be considered is how to use the knowledge of x_t to improve the accuracy of the estimate based on y_t alone.

The solution suggested by Harvey and Chung (2000) is to model the bivariate series $(y_t, x_t)'$ by the structural time series model

$$\begin{aligned} \begin{pmatrix} y_t \\ x_t \end{pmatrix} &= \mu_t + \varepsilon_t, \qquad & \varepsilon_t &\sim \mathrm{N}(0, \Sigma_\varepsilon), \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, \qquad & \xi_t &\sim \mathrm{N}(0, \Sigma_\xi), \\ \nu_{t+1} &= \nu_t + \zeta_t, \qquad & \zeta_t &\sim \mathrm{N}(0, \Sigma_\zeta), \end{aligned} \tag{3.49}$$

for $t = 1, \ldots, n$. Here, μ_t, ε_t, ν_t, ξ_t and ζ_t are 2×1 vectors and Σ_ε, Σ_ξ and Σ_ζ are 2×2 variance matrices. Seasonals can also be incorporated. Many complications are involved in implementing the analysis based on this model, particularly those arising from design features of the survey such as overlapping samples. A point of particular interest is that the claimant count x_t is available one month ahead of the survey value y_t. This extra value of x_t can be easily and efficiently made use of by the missing observations technique discussed in Section 4.10. For a discussion of the details we refer the reader to Harvey and Chung (2000).

This is not the only way in which the information available in x_t can be utilised. For example, in the published discussion of the paper, Durbin (2000a) suggested two further possibilities, the first of which is to model the series $y_t - x_t$ by a structural time series model of one of the forms considered in Section 3.2; unemployment level could then be estimated by $\hat{\mu}_t + x_t$ where $\hat{\mu}_t$ is the forecast of the trend μ_t in the model using information up to time $t - 1$, while the month-to-month change could be estimated by $\hat{\nu}_t + x_t - x_{t-1}$ where $\hat{\nu}_t$ is the forecast of the slope ν_t. Alternatively, x_t could be incorporated as an explanatory variable into an appropriate form of model (3.14) with coefficient β_j replaced by β_{jt} which varies over time according to (3.15). In an obvious notation, trend and change would then be estimated by $\hat{\mu}_t + \hat{\beta}_t x_t$ and $\hat{\nu}_t + \hat{\beta}_t x_t - \hat{\beta}_{t-1} x_{t-1}$.

3.11 Exercises

3.11.1

Consider the autoregressive moving average, with constant, plus noise model

$$y_t = y_t^* + \varepsilon_t, \qquad y_t^* = \mu + \phi_1 y_{t-1}^* + \phi_2 y_{t-2}^* + \zeta_t + \theta_1 \zeta_{t-1},$$

for $t = 1, \ldots, n$, where μ is an unknown constant. The disturbances $\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ and $\zeta_t \sim \mathrm{N}(0, \sigma_\zeta^2)$ are mutually and serially independent at all times and lags. The autoregressive coefficients ϕ_1 and ϕ_2 are restricted such that y_t^* is a stationary process and with moving average coefficient $0 < \theta_1 < 1$. Represent this model in the state space form (3.1); define the state vector α_t and its initial condition.

3.11.2

The local linear trend model with a smooth slope equation is given by

$$y_t = \mu_t + \varepsilon_t, \qquad \mu_{t+1} = \mu_t + \beta_t + \eta_t, \qquad \Delta^r \beta_t = \zeta_t,$$

for $t = 1, \ldots, n$ and some positive integer r, where the normally distributed disturbances ε_t, η_t and ζ_t are mutually and serially independent at all times and lags. Express the trend model for $r = 3$ in the state space form (3.1); define the state vector α_t and its initial condition.

When a stationary slope component is requested, we consider

$$(1 - \phi L)^r \beta_t = \zeta_t,$$

where $0 < \phi < 1$ is an autoregressive coefficient and L is the lag operator. Express the trend model with such a stationary slope for $r = 2$ in the state space form (3.1); define the state state vector α_t and its initial condition.

3.11.3

The exponential smoothing method of forecasting of Section 3.5 can also be expressed by

$$\hat{y}_{t+1} = \hat{y}_t + \lambda(y_t - \hat{y}_t), \qquad t = 1, \ldots, n.$$

where the new one-step ahead forecast is the old one plus an adjustment for the error that occurred in the last forecast. The steady state solution of the Kalman filter applied to the local level model (2.3) in Section 2.11 can be partly represented by exponentional smoothing. Show this relationship and develop an expression of λ in terms of $q = \sigma_\eta^2 \,/\, \sigma_\varepsilon^2$.

3.11.4

Consider the state space model (3.1) extended with regression effects

$$y_t = X_t \beta + Z_t \alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = W_t \beta + T_t \alpha_t + R_t \eta_t,$$

for $t = 1, \ldots, n$, where X_t and W_t are fixed matrices that (partly) consist of exogenous variables and β is a vector of regression coefficient; the matrices have appropriate dimensions. Show that this state space model can be expressed as

$$y_t = X_t^* \beta + Z_t \alpha_t^* + \varepsilon_t, \qquad \alpha_{t+1}^* = T_t \alpha_t^* + R_t \eta_t,$$

for $t = 1, \ldots, n$. Give an expression of X_t^* in terms of X_t, W_t and the other system matrices.

4 Filtering, smoothing and forecasting

4.1 Introduction

In this chapter and the following three chapters we provide a general treatment from both classical and Bayesian perspectives of the linear Gaussian state space model (3.1). The observations y_t will be treated as multivariate. For much of the theory, the development is a straightforward extension to the general case of the treatment of the simple local level model in Chapter 2. We also consider linear unbiased estimates in the non-normal case.

In Section 4.2 we present some elementary results in multivariate regression theory which provide the foundation for our treatment of Kalman filtering and smoothing later in the chapter. We begin by considering a pair of jointly distributed random vectors x and y. Assuming that their joint distribution is normal, we show in Lemma 1 that the conditional distribution of x given y is normal and we derive its mean vector and variance matrix. We shall show in Section 4.3 that these results lead directly to the Kalman filter. For workers who do not wish to assume normality we derive in Lemma 2 the minimum variance linear unbiased estimate of x given y. For those who prefer a Bayesian approach we derive in Lemma 3, under the assumption of normality, the posterior density of x given an observed value of y. Finally in Lemma 4, while retaining the Bayesian approach, we drop the assumption of normality and derive a quasi-posterior density of x given y, with a mean vector which is linear in y and which has minimum variance matrix.

All four lemmas can be regarded as representing in appropriate senses the regression of x on y. For this reason in all cases the mean vectors and variance matrices are the same. We shall use these lemmas to derive the Kalman filter and smoother in Sections 4.3 and 4.4. Because the mean vectors and variance matrices are the same, we need only use one of the four lemmas to derive the results that we need; the results so obtained then remain valid under the conditions assumed under the other three lemmas.

Denote the set of observations $y_1, \ldots, y_t$ by Y_t. In Section 4.3 we will derive the Kalman filter, which is a recursion for calculating $a_{t|t} = \mathrm{E}(\alpha_t | Y_t)$, $a_{t+1} = \mathrm{E}(\alpha_{t+1} | Y_t)$, $P_{t|t} = \mathrm{Var}(\alpha_t | Y_t)$ and $P_{t+1} = \mathrm{Var}(\alpha_{t+1} | Y_t)$ given a_t and P_t. The derivation requires only elementary properties of multivariate regression theory derived in Lemmas 1 to 4. We also investigate some properties of state estimation errors and one-step ahead forecast errors. In Section 4.4 we use the output of the Kalman filter and the properties of forecast errors to obtain recursions for smoothing the series, that is, calculating the conditional mean and variance matrix of α_t, for $t = 1, \dots, n, n+1$, given all the observations $y_1, \dots, y_n$. Estimates of the disturbance vectors ε_t and η_t and their error variance matrices given all the data are derived in Section 4.5. Covariance matrices of smoothed estimators are considered in Section 4.7. The weights associated with filtered and smoothed estimates of functions of the state and disturbance vectors are discussed in Section 4.8. Section 4.9 describes how to generate random samples for purposes of simulation from the smoothed densities of the state and disturbance vectors given the observations. The problem of missing observations is considered in Section 4.10 where we show that with the state space approach the problem is easily dealt with by means of simple modifications of the Kalman filter and the smoothing recursions. Section 4.11 shows that forecasts of observations and state can be obtained simply by treating future observations as missing values; these results are of special significance in view of the importance of forecasting in much practical time series work. A comment on varying dimensions of the observation vector is given in Section 4.12. Finally, in Section 4.13 we consider a general matrix formulation of the state space model.

4.2 Basic results in multivariate regression theory

In this section we present some basic results in elementary multivariate regression theory that we shall use for the development of the theory for the linear Gaussian state space model (3.1) and its non-Gaussian version with $\varepsilon_t \sim (0, H_t)$ and $\eta_t \sim (0, Q_t)$. We shall present the results in a general form before embarking on the state space theory because we shall need to apply them in a variety of different contexts and it is preferable to prove them once only in general rather than to produce a series of similar *ad hoc* proofs tailored to specific situations. A further point is that this form of presentation exposes the intrinsically simple nature of the mathematical theory underlying the state space approach to time series analysis. Readers who are prepared to take for granted the results in Lemmas 1 to 4 below can skip the proofs and go straight to Section 4.3.

Suppose that x and y are jointly normally distributed random vectors with

$$\mathrm{E}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \qquad \mathrm{Var}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \Sigma_{yy} \end{bmatrix}, \tag{4.1}$$

where Σ_{yy} is assumed to be a nonsingular matrix.

Lemma 1 *The conditional distribution of x given y is normal with mean vector*

$$\mathrm{E}(x|y) = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \tag{4.2}$$

and variance matrix

$$\mathrm{Var}(x|y) = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}'. \tag{4.3}$$

Proof. Let $z = x - \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$. Since the transformation from (x, y) to (y, z) is linear and (x, y) is normally distributed, the joint distribution of y and z is normal. We have

$$\begin{aligned}
\mathrm{E}(z) &= \mu_x \\
\mathrm{Var}(z) &= \mathrm{E}\left[(z - \mu_x)(z - \mu_x)'\right] \\
&= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}', && (4.4) \\
\mathrm{Cov}(y, z) &= \mathrm{E}\left[y(z - \mu_x)'\right] \\
&= \mathrm{E}\left[y(x - \mu_x)' - y(y - \mu_y)'\Sigma_{yy}^{-1}\Sigma_{xy}'\right] \\
&= 0. && (4.5)
\end{aligned}$$

Using the result that if two vectors are normal and uncorrelated they are independent, we infer from (4.5) that z is distributed independently of y. Since the distribution of z does not depend on y its conditional distribution given y is the same as its unconditional distribution, that is, it is normal with mean vector μ_x and variance matrix (4.4) which is the same as (4.3). Since $z = x - \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$, it follows that the conditional distribution of x given y is normal with mean vector (4.2) and variance matrix (4.3). □

Formulae (4.2) and (4.3) are well known in regression theory. An early proof in a state space context is given in Åström (1970, Chapter 7, Theorem 3.2). The proof given here is based on the treatment given by Rao (1973, §8a.2(v)). A partially similar proof is given by Anderson (2003, Theorem 2.5.1). A quite different proof in a state space context is given by Anderson and Moore (1979, Example 3.2) which is repeated by Harvey (1989, Appendix to Chapter 3); some details of this proof are given in Exercise 4.14.1.

We can regard Lemma 1 as representing the regression of x on y in a multivariate normal distribution. It should be noted that Lemma 1 remains valid when Σ_{yy} is singular if the symbol Σ_{yy}^{-1} is interpreted as a generalised inverse; see the treatment in Rao (1973). Åström (1970) pointed out that if the distribution of (x, y) is singular we can always derive a nonsingular distribution by making a projection on the hyperplanes where the mass is concentrated. The fact that the conditional variance $\mathrm{Var}(x|y)$ given by (4.3) does not depend on y is a property special to the multivariate normal distribution and does not generally hold for other distributions.

We now consider the estimation of x when x is unknown and y is known, as for example when y is an observed vector. Under the assumptions of Lemma 1 we take as our estimate of x the conditional expectation $\widehat{x} = \mathrm{E}(x|y)$, that is,

$$\widehat{x} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y). \tag{4.6}$$

This has estimation error $\widehat{x} - x$ so $\widehat{x}$ is conditionally unbiased in the sense that $\mathrm{E}(\widehat{x} - x|y) = \widehat{x} - \mathrm{E}(x|y) = 0$. It is also obviously unconditionally unbiased in the sense that $\mathrm{E}(\widehat{x} - x) = 0$. The unconditional error variance matrix of $\widehat{x}$ is

$$\mathrm{Var}(\widehat{x}-x)=\mathrm{Var}\left[\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)-(x-\mu_x)\right] \;=\; \Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}'. \quad (4.7)$$

Expressions (4.6) and (4.7) are, of course, the same as (4.2) and (4.3) respectively.

We now consider the estimation of x given y when the assumption that (x, y) is normally distributed is dropped. We assume that the other assumptions of Lemma 1 are retained. Let us restrict our attention to estimates $\bar{x}$ that are linear in the elements of y, that is, we shall take

$$\bar{x}=\beta+\gamma y,$$

where β is a fixed vector and γ is a fixed matrix of appropriate dimensions. The estimation error is $\bar{x}-x$. If $\mathrm{E}(\bar{x}-x)=0$, we say that $\bar{x}$ is a *linear unbiased estimate* (LUE) of x given y. If there is a particular value x^* of $\bar{x}$ such that

$$\mathrm{Var}(\bar{x}-x)-\mathrm{Var}(x^*-x),$$

is non-negative definite for all LUEs $\bar{x}$ we say that x^* is a *minimum variance linear unbiased estimate* (MVLUE) of x given y. Note that the mean vectors and variance matrices here are unconditional and not conditional given y as were considered in Lemma 1. An MVLUE for the non-normal case is given by the following lemma.

Lemma 2 *Whether (x,y) is normally distributed or not, the estimate $\widehat{x}$ defined by* (4.6) *is a MVLUE of x given y and its error variance matrix is given by* (4.7).

Proof. Since $\bar{x}$ is an LUE, we have

$$\begin{aligned}\mathrm{E}(\bar{x}-x)&=\mathrm{E}(\beta+\gamma y-x)\\&=\beta+\gamma\mu_y-\mu_x \;=\; 0.\end{aligned}$$

It follows that $\beta=\mu_x-\gamma\mu_y$ and therefore

$$\bar{x}=\mu_x+\gamma(y-\mu_y). \quad (4.8)$$

Thus

$$\begin{aligned}\mathrm{Var}(\bar{x}-x)&=\mathrm{Var}\left[\mu_x+\gamma(y-\mu_y)-x\right]\\&=\mathrm{Var}\left[\gamma(y-\mu_y)-(x-\mu_x)\right]\\&=\gamma\Sigma_{yy}\gamma'-\gamma\Sigma_{xy}'-\Sigma_{xy}\gamma'+\Sigma_{xx}\\&=\mathrm{Var}\left[(\gamma-\Sigma_{xy}\Sigma_{yy}^{-1})y\right]+\Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}'. \quad (4.9)\end{aligned}$$

Let $\widehat{x}$ be the value of $\bar{x}$ obtained by putting $\gamma=\Sigma_{xy}\Sigma_{yy}^{-1}$ in (4.8). Then $\widehat{x}=\mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)$ and from (4.9), it follows that

$$\mathrm{Var}(\widehat{x}-x)=\Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}'.$$

We can therefore rewrite (4.9) in the form

$$\mathrm{Var}(\bar{x} - x) = \mathrm{Var}\left[(\gamma - \Sigma_{xy}\Sigma_{yy}^{-1})y\right] + \mathrm{Var}(\widehat{x} - x), \tag{4.10}$$

which holds for all LUEs $\bar{x}$. Since $\mathrm{Var}\left[(\gamma - \Sigma_{xy}\Sigma_{yy}^{-1})y\right]$ is non-negative definite the lemma is proved. □

The MVLUE property of the vector estimate $\widehat{x}$ implies that arbitrary linear functions of elements of $\widehat{x}$ are minimum variance linear unbiased estimates of the corresponding linear functions of the elements of x. Lemma 2 can be regarded as an analogue for multivariate distributions of the Gauss–Markov theorem for least squares regression of a dependent variable on fixed regressors. For a treatment of the Gauss–Markov theorem, see, for example, Davidson and MacKinnon (1993, Chapter 3). Lemma 2 is proved in the special context of Kalman filtering by Duncan and Horn (1972) and by Anderson and Moore (1979, §3.2). However, their treatments lack the brevity and generality of Lemma 2 and its proof.

Lemma 2 is highly significant for workers who prefer not to assume normality as the basis for the analysis of time series on the grounds that many real time series have distributions that appear to be far from normal; however, the MVLUE criterion is regarded as acceptable as a basis for analysis by many of these workers. We will show later in the book that many important results in state space analysis such as Kalman filtering and smoothing, missing observation analysis and forecasting can be obtained by using Lemma 1; Lemma 2 shows that these results also satisfy the MVLUE criterion. A variant of Lemma 2 is to formulate it in terms of minimum mean square error matrix rather than minimum variance unbiasedness; this variant is dealt with in Exercise 4.14.4.

Other workers prefer to treat inference problems in state space time series analysis from a Bayesian point of view instead of from the classical standpoint for which Lemmas 1 and 2 are appropriate. We therefore consider basic results in multivariate regression theory that will lead us to a Bayesian treatment of the linear Gaussian state space model.

Suppose that x is a parameter vector with *prior density* $p(x)$ and that y is an observational vector with density $p(y)$ and conditional density $p(y|x)$. Suppose further that the joint density of x and y is the multivariate normal density $p(x, y)$. Then the *posterior density* of x given y is

$$p(x|y) = \frac{p(x, y)}{p(y)} = \frac{p(x)p(y|x)}{p(y)}. \tag{4.11}$$

We shall use the same notation as in (4.1) for the first and second moments of x and y. The equation (4.11) is a form of Bayes Theorem.

Lemma 3 *The posterior density of x given y is normal with posterior mean vector* (4.2) *and posterior variance matrix* (4.3).

Proof. Using (4.11), the proof follows immediately from Lemma 1. □

A general Bayesian treatment of state space time series analysis is given by West and Harrison (1997, §17.2.2) in which an explicit proof of our Lemma 3 is given. Their results emerge in a different form from our (4.2) and (4.3) which, as they point out, can be converted to ours by an algebraical identity; see Exercise 4.14.5 for details.

We now drop the assumption of normality and derive a Bayesian-type analogue of Lemma 2. Let us introduce a broader concept of a posterior density than the traditional one by using the term 'posterior density' to refer to *any* density of x given y and not solely to the conditional density of x given y. Let us consider posterior densities whose mean $\bar{x}$ given y is linear in the elements of y, that is, we take $\bar{x} = \beta + \gamma y$ where β is a fixed vector and γ is a fixed matrix; we say that $\bar{x}$ is a *linear posterior mean.* If there is a particular value x^* of $\bar{x}$ such that

$$\mathrm{Var}(\bar{x} - x) - \mathrm{Var}(x^* - x)$$

is non-negative definite for all linear posterior means $\bar{x}$, we say that x^* is a *minimum variance linear posterior mean estimate* (MVLPME) of x given y.

Lemma 4 *The linear posterior mean $\widehat{x}$ defined by* (4.6) *is a MVLPME and its error variance matrix is given by* (4.7).

Proof. Taking expectations with respect to density $p(y)$ we have

$$\mathrm{E}(\bar{x}) = \mu_x = \mathrm{E}(\beta + \gamma y) = \beta + \gamma \mu_y,$$

from which it follows that $\beta = \mu_x - \gamma\mu_y$ and hence that (4.8) holds. Let $\widehat{x}$ be the value of $\bar{x}$ obtained by putting $\gamma = \Sigma_{xy}\Sigma_{yy}^{-1}$ in (4.8). It follows as in the proof of Lemma 2 that (4.10) applies so the lemma is proved. □

The four lemmas in this section have an important common property. Although each lemma starts from a different criterion, they all finish up with distributions which have the same mean vector (4.2) and the same variance matrix (4.3). The significance of this result is that formulae for the Kalman filter, its associated smoother and related results throughout Part I of the book are exactly the same whether an individual worker wishes to start from a criterion of classical inference, minimum variance linear unbiased estimation or Bayesian inference.

4.3 Filtering

4.3.1 Derivation of the Kalman filter

For convenience we restate the linear Gaussian state space model (3.1) here as

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{N}(0, H_t), \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t, & \eta_t &\sim \mathrm{N}(0, Q_t), & t &= 1, \dots, n, \\ & & \alpha_1 &\sim \mathrm{N}(a_1, P_1), \end{aligned} \tag{4.12}$$

where details are given below (3.1). At various points we shall drop the normality assumptions in (4.12). Let Y_{t-1} denote the set of past observations $y_1, \dots, y_{t-1}$ for $t = 2, 3, \dots$ while Y_0 indicates that there is no prior observation before $t = 1$. In our treatments below, we will define Y_t by the vector $(y_1', \dots, y_t')'$. Starting at $t = 1$ in (4.12) and building up the distributions of α_t and y_t recursively, it is easy to show that $p(y_t|\alpha_1, \dots, \alpha_t, Y_{t-1}) = p(y_t|\alpha_t)$ and $p(\alpha_{t+1}|\alpha_1, \dots, \alpha_t, Y_t) = p(\alpha_{t+1}|\alpha_t)$. In Table 4.1 we give the dimensions of the vectors and matrices of the state space model.

In this section we derive the Kalman filter for model (4.12) for the case where the initial state α_1 is $\mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known. We shall base the derivation on classical inference using Lemma 1. It follows from Lemmas 2 to 4 that the basic results are also valid for minimum variance linear unbiased estimation and for Bayesian-type inference with or without the normality assumption. Returning to the assumption of normality, our object is to obtain the conditional distributions of α_t and α_{t+1} given Y_t for $t = 1, \dots, n$. Let $a_{t|t} = \mathrm{E}(\alpha_t|Y_t)$, $a_{t+1} = \mathrm{E}(\alpha_{t+1}|Y_t)$, $P_{t|t} = \mathrm{Var}(\alpha_t|Y_t)$ and $P_{t+1} = \mathrm{Var}(\alpha_{t+1}|Y_t)$. Since all distributions are normal, it follows from Lemma 1 that conditional distributions of subsets of variables given other subsets of variables are also normal; the distributions of α_t given Y_t and α_{t+1} given Y_t are therefore given by $\mathrm{N}(a_{t|t}, P_{t|t})$ and $\mathrm{N}(a_{t+1}, P_{t+1})$. We proceed inductively; starting with $\mathrm{N}(a_t, P_t)$, the distribution of α_t given Y_{t-1}, we show how to calculate $a_{t|t}$, a_{t+1}, $P_{t|t}$ and P_{t+1} from a_t and P_t recursively for $t = 1, \dots, n$.

Let

$$v_t = y_t - \mathrm{E}(y_t|Y_{t-1}) = y_t - \mathrm{E}(Z_t\alpha_t + \varepsilon_t|Y_{t-1}) = y_t - Z_t a_t. \tag{4.13}$$

Table 4.1 Dimensions of state space model (4.12).

Vector		Matrix	
y_t	$p \times 1$	Z_t	$p \times m$
α_t	$m \times 1$	T_t	$m \times m$
ε_t	$p \times 1$	H_t	$p \times p$
η_t	$r \times 1$	R_t	$m \times r$
		Q_t	$r \times r$
a_1	$m \times 1$	P_1	$m \times m$

Thus v_t is the one-step ahead forecast error of y_t given Y_{t-1}. When Y_{t-1} and v_t are fixed then Y_t is fixed and vice versa. Thus $\mathrm{E}(\alpha_t|Y_t) = \mathrm{E}(\alpha_t|Y_{t-1}, v_t)$. But $\mathrm{E}(v_t|Y_{t-1}) = \mathrm{E}(y_t - Z_t a_t|Y_{t-1}) = \mathrm{E}(Z_t\alpha_t + \varepsilon_t - Z_t a_t|Y_{t-1}) = 0$. Consequently, $\mathrm{E}(v_t) = 0$ and $\mathrm{Cov}(y_j, v_t) = \mathrm{E}[y_j \mathrm{E}(v_t|Y_{t-1})'] = 0$ for $j = 1, \dots, t-1$. Also,

$$\begin{aligned} a_{t|t} &= \mathrm{E}(\alpha_t|Y_t) = \mathrm{E}(\alpha_t|Y_{t-1}, v_t), \\ a_{t+1} &= \mathrm{E}(\alpha_{t+1}|Y_t) = \mathrm{E}(\alpha_{t+1}|Y_{t-1}, v_t). \end{aligned}$$

Now apply Lemma 1 in Section 4.2 to the conditional joint distribution of α_t and v_t given Y_{t-1}, taking x and y in Lemma 1 as α_t and v_t here. This gives

$$a_{t|t} = \mathrm{E}(\alpha_t|Y_{t-1}) + \mathrm{Cov}(\alpha_t, v_t)[\mathrm{Var}(v_t)]^{-1} v_t, \tag{4.14}$$

where Cov and Var refer to covariance and variance in the conditional joint distributions of α_t and v_t given Y_{t-1}. Here, $\mathrm{E}(\alpha_t|Y_{t-1}) = a_t$ by definition of a_t and

$$\begin{aligned} \mathrm{Cov}(\alpha_t, v_t) &= \mathrm{E}\left[\alpha_t \left(Z_t\alpha_t + \varepsilon_t - Z_t a_t\right)' |Y_{t-1}\right] \\ &= \mathrm{E}\left[\alpha_t(\alpha_t - a_t)' Z_t' |Y_{t-1}\right] = P_t Z_t', \end{aligned} \tag{4.15}$$

by definition of P_t. Let

$$F_t = \mathrm{Var}(v_t|Y_{t-1}) = \mathrm{Var}(Z_t\alpha_t + \varepsilon_t - Z_t a_t|Y_{t-1}) = Z_t P_t Z_t' + H_t. \tag{4.16}$$

Then

$$a_{t|t} = a_t + P_t Z_t' F_t^{-1} v_t. \tag{4.17}$$

By (4.3) of Lemma 1 in Section 4.2 we have

$$\begin{aligned} P_{t|t} = \mathrm{Var}(\alpha_t|Y_t) &= \mathrm{Var}(\alpha_t|Y_{t-1}, v_t) \\ &= \mathrm{Var}(\alpha_t|Y_{t-1}) - \mathrm{Cov}(\alpha_t, v_t)[\mathrm{Var}(v_t)]^{-1} \mathrm{Cov}(\alpha_t, v_t)' \\ &= P_t - P_t Z_t' F_t^{-1} Z_t P_t. \end{aligned} \tag{4.18}$$

We assume that F_t is nonsingular; this assumption is normally valid in well-formulated models, but in any case it is relaxed in Section 6.4. Relations (4.17) and (4.18) are sometimes called the *updating step* of the Kalman filter.

We now develop recursions for a_{t+1} and P_{t+1}. Since $\alpha_{t+1} = T_t\alpha_t + R_t\eta_t$, we have

$$\begin{aligned} a_{t+1} &= \mathrm{E}(T_t\alpha_t + R_t\eta_t|Y_t) \\ &= T_t\, \mathrm{E}(\alpha_t|Y_t), \end{aligned} \tag{4.19}$$

$$\begin{aligned} P_{t+1} &= \mathrm{Var}(T_t\alpha_t + R_t\eta_t|Y_t) \\ &= T_t\ \mathrm{Var}(\alpha_t|Y_t) T_t' + R_t Q_t R_t', \end{aligned} \tag{4.20}$$

for $t = 1, \dots, n$.

Substituting (4.17) into (4.19) gives

$$\begin{aligned} a_{t+1} &= T_t a_{t|t} \\ &= T_t a_t + K_t v_t, \qquad t = 1, \dots, n, \end{aligned} \tag{4.21}$$

where

$$K_t = T_t P_t Z_t' F_t^{-1}. \tag{4.22}$$

The matrix K_t is referred to as the *Kalman gain*. We observe that a_{t+1} has been obtained as a linear function of the previous value a_t and the forecast error v_t of y_t given Y_{t-1}. Substituting from (4.18) and (4.22) in (4.20) gives

$$P_{t+1} = T_t P_t (T_t - K_t Z_t)' + R_t Q_t R_t', \qquad t = 1, \dots, n. \tag{4.23}$$

Relations (4.21) and (4.23) are sometimes called the *prediction step* of the Kalman filter.

The recursions (4.17), (4.21), (4.18) and (4.23) constitute the celebrated Kalman filter for model (4.12). They enable us to update our knowledge of the system each time a new observation comes in. It is noteworthy that we have derived these recursions by simple applications of the standard results of multivariate normal regression theory contained in Lemma 1. The key advantage of the recursions is that we do not have to invert a $(pt \times pt)$ matrix to fit the model each time the tth observation comes in for $t = 1, \dots, n$; we only have to invert the $(p \times p)$ matrix F_t and p is generally much smaller than n; indeed, in the most important case in practice, $p = 1$. Although relations (4.17), (4.21), (4.18) and (4.23) constitute the forms in which the multivariate Kalman filter recursions are usually presented, we shall show in Section 6.4 that variants of them in which elements of the observational vector y_t are brought in one at a time, rather than the entire vector y_t, are in general computationally superior.

We infer from Lemma 2 that when the observations are not normally distributed and we restrict attention to estimates which are linear in y_t and unbiased, and also when matrices Z_t and T_t do not depend on previous y_t's, then under appropriate assumptions the values of $a_{t|t}$ and a_{t+1} given by the filter minimise the variance matrices of the estimates of α_t and α_{t+1} given Y_t. These considerations emphasise the point that although our results are obtained under the assumption of normality, they have a wider validity in the sense of minimum variance linear unbiased estimation when the variables involved are not normally distributed. It follows from the discussion just after the proof of Lemma 2 that the estimates are also minimum error variance linear estimates.

From the standpoint of Bayesian inference we note that, on the assumption of normality, Lemma 3 implies that the posterior densities of α_t and α_{t+1} given Y_t are normal with mean vectors (4.17) and (4.21) and variance matrices (4.18) and (4.23), respectively. We therefore do not need to provide a separate Bayesian derivation of the Kalman filter. If the assumption of normality is dropped, Lemma 4 demonstrates that the Kalman filter, as we have derived it, provides quasi-posterior mean vectors and variance matrices with minimum variance linear unbiased interpretations.

4.3.2 Kalman filter recursion

For convenience we collect together the filtering equations

$$\begin{aligned} v_t &= y_t - Z_t a_t, & F_t &= Z_t P_t Z_t' + H_t, \\ a_{t|t} &= a_t + P_t Z_t' F_t^{-1} v_t, & P_{t|t} &= P_t - P_t Z_t' F_t^{-1} Z_t P_t, \\ a_{t+1} &= T_t a_t + K_t v_t, & P_{t+1} &= T_t P_t (T_t - K_t Z_t)' + R_t Q_t R_t', \end{aligned} \tag{4.24}$$

for $t = 1, \ldots, n$, where $K_t = T_t P_t Z_t' F_t^{-1}$ with a_1 and P_1 as the mean vector and variance matrix of the initial state vector α_1. The recursion (4.24) is called the *Kalman filter*. Once $a_{t|t}$ and $P_{t|t}$ are computed, it suffices to adopt the relations

$$a_{t+1} = T_t a_{t|t}, \qquad P_{t+1} = T_t P_{t|t} T_t' + R_t Q_t R_t',$$

for predicting the state vector α_{t+1} and its variance matrix at time t. In Table 4.2 we give the dimensions of the vectors and matrices of the Kalman filter equations.

4.3.3 Kalman filter for models with mean adjustments

It is sometimes convenient to include mean adjustments in the state space model (4.12) giving the form

$$\begin{aligned} y_t &= Z_t \alpha_t + d_t + \varepsilon_t, & \varepsilon_t &\sim N(0, H_t), \\ \alpha_{t+1} &= T_t \alpha_t + c_t + R_t \eta_t, & \eta_t &\sim N(0, Q_t), \\ && \alpha_1 &\sim N(a_1, P_1), \end{aligned} \tag{4.25}$$

where $p \times 1$ vector d_t and $m \times 1$ vector c_t are known and may change over time. Indeed, Harvey (1989) employs (4.25) as the basis for the treatment of the linear Gaussian state space model. While the simpler model (4.12) is adequate for most purposes, it is worth while presenting the Kalman filter for model (4.25) explicitly for occasional use.

Table 4.2 Dimensions of Kalman filter.

Vector		Matrix	
v_t	$p \times 1$	F_t	$p \times p$
		K_t	$m \times p$
a_t	$m \times 1$	P_t	$m \times m$
$a_{t\|t}$	$m \times 1$	$P_{t\|t}$	$m \times m$

Defining $a_t = \mathrm{E}(\alpha_t|Y_{t-1})$ and $P_t = \mathrm{Var}(\alpha_t|Y_{t-1})$ as before and assuming that d_t can depend on Y_{t-1} and c_t can depend on Y_t, the Kalman filter for (4.25) takes the form

$$\begin{aligned} v_t &= y_t - Z_t a_t - d_t, & F_t &= Z_t P_t Z_t' + H_t, \\ a_{t|t} &= a_t + P_t Z_t' F_t^{-1} v_t, & P_{t|t} &= P_t - P_t Z_t' F_t^{-1} Z_t P_t, \\ a_{t+1} &= T_t a_{t|t} + c_t, & P_{t+1} &= T_t P_{t|t} T_t' + R_t Q_t R_t', \end{aligned} \tag{4.26}$$

for $t = 1, \ldots, n$. The reader can easily verify this result by going through the argument leading from (4.19) to (4.23) step by step for model (4.25) in place of model (4.12).

4.3.4 Steady state

When dealing with a time-invariant state space model in which the system matrices Z_t, H_t, T_t, R_t, and Q_t are constant over time, the Kalman recursion for P_{t+1} converges to a constant matrix $\bar{P}$ which is the solution to the matrix equation

$$\bar{P} = T\bar{P}T' - T\bar{P}Z'\bar{F}^{-1}Z\bar{P}T' + RQR',$$

where $\bar{F} = Z\bar{P}Z' + H$. The solution that is reached after convergence to $\bar{P}$ is referred to as the *steady state solution* of the Kalman filter. Use of the steady state after convergence leads to considerable computational savings because the recursive computations for F_t, K_t, $P_{t|t}$ and P_{t+1} are no longer required.

4.3.5 State estimation errors and forecast errors

Define the *state estimation error* as

$$x_t = \alpha_t - a_t, \quad \text{with} \quad \mathrm{Var}(x_t) = P_t, \tag{4.27}$$

as for the local level model in Subsection 2.3.2. We now investigate how these errors are related to each other and to the one-step ahead forecast errors $v_t = y_t - \mathrm{E}(y_t|Y_{t-1}) = y_t - Z_t a_t$. Since v_t is the part of y_t that cannot be predicted from the past, the v_t's are sometimes referred to as *innovations*. It follows immediately from the Kalman filter relations and the definition of x_t that

$$\begin{aligned} v_t &= y_t - Z_t a_t \\ &= Z_t \alpha_t + \varepsilon_t - Z_t a_t \\ &= Z_t x_t + \varepsilon_t, \end{aligned} \tag{4.28}$$

and

$$\begin{aligned} x_{t+1} &= \alpha_{t+1} - a_{t+1} \\ &= T_t \alpha_t + R_t \eta_t - T_t a_t - K_t v_t \\ &= T_t x_t + R_t \eta_t - K_t Z_t x_t - K_t \varepsilon_t \\ &= L_t x_t + R_t \eta_t - K_t \varepsilon_t, \end{aligned} \tag{4.29}$$

where $K_t = T_t P_t Z_t' F_t^{-1}$ and $L_t = T_t - K_t Z_t$; these recursions are similar to (2.31) for the local level model. Analogously to the state space relations

$$y_t = Z_t \alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = T_t \alpha_t + R_t \eta_t,$$

we obtain the *innovation analogue* of the state space model, that is,

$$v_t = Z_t x_t + \varepsilon_t, \qquad x_{t+1} = L_t x_t + R_t \eta_t - K_t \varepsilon_t, \tag{4.30}$$

with $x_1 = \alpha_1 - a_1$, for $t = 1, \ldots, n$. The recursion for P_{t+1} can be derived more easily than in Subsection 4.3.1 by the steps

$$\begin{aligned} P_{t+1} &= \mathrm{Var}(x_{t+1}) = \mathrm{E}[(\alpha_{t+1} - a_{t+1}) x_{t+1}'] \\ &= \mathrm{E}(\alpha_{t+1} x_{t+1}') \\ &= \mathrm{E}[(T_t \alpha_t + R_t \eta_t)(L_t x_t + R_t \eta_t - K_t \varepsilon_t)'] \\ &= T_t P_t L_t' + R_t Q_t R_t', \end{aligned}$$

since $\mathrm{Cov}(x_t, \eta_t) = 0$. Relations (4.30) will be used for deriving the smoothing recursions in the next section.

We finally show that the one-step ahead forecast errors are independent of each other using the same arguments as in Subsection 2.3.1. The joint density of the observational vectors $y_1, \ldots, y_n$ is

$$p(y_1, \ldots, y_n) = p(y_1) \prod_{t=2}^{n} p(y_t | Y_{t-1}).$$

Transforming from y_t to $v_t = y_t - Z_t a_t$ we have

$$p(v_1, \ldots, v_n) = \prod_{t=1}^{n} p(v_t),$$

since $p(y_1) = p(v_1)$ and the Jacobian of the transformation is unity because each v_t is y_t minus a linear function of $y_1, \ldots, y_{t-1}$ for $t = 2, \ldots, n$. Consequently $v_1, \ldots, v_n$ are independent of each other, from which it also follows that $v_t, \ldots, v_n$ are independent of Y_{t-1}.

4.4 State smoothing

4.4.1 Introduction

We now derive the conditional density of α_t given the entire series $y_1, \ldots, y_n$ for $t = 1, \ldots, n$. We do so by assuming normality and using Lemma 1, noting from Lemmas 2 to 4 that the mean vectors and variance matrices we obtain are valid without the normality assumption in the minimum variance linear unbiased sense and are also valid for Bayesian analyses.

We shall calculate the conditional mean $\hat{\alpha}_t = \mathrm{E}(\alpha_t|Y_n)$ and the conditional variance matrix $V_t = \mathrm{Var}(\alpha_t|Y_n)$ for $t = 1, \ldots, n$. Our approach is to construct recursions for $\hat{\alpha}_t$ and V_t on the assumption that $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known, deferring consideration of the case a_1 and P_1 unknown until Chapter 5. The operation of calculating $\hat{\alpha}_t$ is called *state smoothing* or just *smoothing*. A conditional mean $\mathrm{E}(\alpha_t|y_t, \ldots, y_s)$ is sometimes called a *fixed-interval smoother* to reflect the fact that it is based on the fixed time interval (t, s). The smoother conditioned on the full sample Y_n as we have just discussed is the most common smoother encountered in practice. Other types of smoothers are the *fixed-point smoother* $\hat{\alpha}_{t|n} = \mathrm{E}(\alpha_t|Y_n)$ for t fixed and $n = t+1, t+2, \ldots$ and the *fixed-lag smoother* $\hat{\alpha}_{n-j|n} = \mathrm{E}(\alpha_{n-j}|Y_n)$ for a fixed positive integer j and $n = j+1, j+2, \ldots$. We shall give formulae for these smoothers in Subsection 4.4.6. The fixed-point and fixed-lag smoothers are important in engineering; see, for example, the treatment in Chapter 7 of Anderson and Moore (1979). However, in this book we shall focus mainly on fixed-interval smoothing and when we refer simply to ‘smoother’ and ‘smoothing’, it is the fixed-interval smoother based on the full sample with which we are concerned.

4.4.2 Smoothed state vector

Take $v_1, \ldots, v_n$ as in Subsection 4.3.1 and denote the vector $(v_t', \ldots, v_n')'$ by $v_{t:n}$; note also that Y_n is fixed when Y_{t-1} and $v_{t:n}$ are fixed. To calculate $\mathrm{E}(\alpha_t|Y_n)$ and $\mathrm{Var}(\alpha_t|Y_n)$ we apply Lemma 1 of Subsection 4.2 to the conditional joint distributions of α_t and $v_{t:n}$ given Y_{t-1}, taking x and y of Lemma 1 as α_t and $v_{t:n}$ here. Using the fact that $v_t, \ldots, v_n$ are independent of Y_{t-1} and of each other with zero means, we therefore have from (4.2),

$$\begin{aligned}\hat{\alpha}_t &= \mathrm{E}(\alpha_t|Y_n) = \mathrm{E}(\alpha_t|Y_{t-1}, v_{t:n}) \\ &= a_t + \sum_{j=t}^{n} \mathrm{Cov}(\alpha_t, v_j) F_j^{-1} v_j, \end{aligned} \tag{4.31}$$

since $\mathrm{E}(\alpha_t|Y_{t-1}) = a_t$ for $t = 1, \ldots, n$, where Cov refers to covariance in the conditional distribution given Y_{t-1} and $F_j = \mathrm{Var}(v_j|Y_{t-1})$. It follows from (4.30) that

$$\begin{aligned}\mathrm{Cov}(\alpha_t, v_j) &= \mathrm{E}(\alpha_t v_j'|Y_{t-1}) \\ &= \mathrm{E}[\alpha_t (Z_j x_j + \varepsilon_j)'|Y_{t-1}] \\ &= \mathrm{E}(\alpha_t x_j'|Y_{t-1}) Z_j', \qquad j = t, \ldots, n. \end{aligned} \tag{4.32}$$

Moreover,

$$\begin{aligned} \mathrm{E}(\alpha_t x_t'|Y_{t-1}) &= \mathrm{E}[\alpha_t(\alpha_t - a_t)|Y_{t-1}] = P_t, \\ \mathrm{E}(\alpha_t x_{t+1}'|Y_{t-1}) &= \mathrm{E}[\alpha_t(L_t x_t + R_t\eta_t - K_t\varepsilon_t)'|Y_{t-1}] = P_t L_t', \\ \mathrm{E}(\alpha_t x_{t+2}'|Y_{t-1}) &= P_t L_t' L_{t+1}', \\ &\;\;\vdots \\ \mathrm{E}(\alpha_t x_n'|Y_{t-1}) &= P_t L_t' \cdots L_{n-1}', \end{aligned} \tag{4.33}$$

using (4.30) repeatedly for $t+1, t+2, \dots$. Note that here and elsewhere we interpret $L_t' \cdots L_{n-1}'$ as I_m when $t = n$ and as L_{n-1}' when $t = n-1$. Substituting into (4.31) gives

$$\begin{aligned} \hat{\alpha}_n &= a_n + P_n Z_n' F_n^{-1} v_n, \\ \hat{\alpha}_{n-1} &= a_{n-1} + P_{n-1} Z_{n-1}' F_{n-1}^{-1} v_{n-1} + P_{n-1} L_n' Z_n' F_n^{-1} v_n, \\ \hat{\alpha}_t &= a_t + P_t Z_t' F_t^{-1} v_t + P_t L_t' Z_{t+1}' F_{t+1}^{-1} v_{t+1} \\ &\quad + \cdots + P_t L_t' \cdots L_{n-1}' Z_n' F_n^{-1} v_n, \end{aligned} \tag{4.34}$$

for $t = n-2, n-3, \dots, 1$. We can therefore express the smoothed state vector as

$$\hat{\alpha}_t = a_t + P_t r_{t-1}, \tag{4.35}$$

where $r_{n-1} = Z_n' F_n^{-1} v_n$, $r_{n-2} = Z_{n-1}' F_{n-1}^{-1} v_{n-1} + L_{n-1}' Z_n' F_n^{-1} v_n$ and

$$r_{t-1} = Z_t' F_t^{-1} v_t + L_t' Z_{t+1}' F_{t+1}^{-1} v_{t+1} + \cdots + L_t' L_{t+1}' \cdots L_{n-1}' Z_n' F_n^{-1} v_n, \tag{4.36}$$

for $t = n-2, n-3, \dots, 1$. The vector r_{t-1} is a weighted sum of innovations v_j occurring after time $t-1$, that is, for $j = t, \dots, n$. The value at time t is

$$r_t = Z_{t+1}' F_{t+1}^{-1} v_{t+1} + L_{t+1}' Z_{t+2}' F_{t+2}^{-1} v_{t+2} + \cdots + L_{t+1}' \cdots L_{n-1}' Z_n' F_n^{-1} v_n; \tag{4.37}$$

also $r_n = 0$ since no innovations are available after time n. Substituting from (4.37) into (4.36) we obtain the backwards recursion

$$r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t, \qquad t = n, \dots, 1, \tag{4.38}$$

with $r_n = 0$.

Collecting these results together gives the recursion for state smoothing,

$$\hat{\alpha}_t = a_t + P_t r_{t-1}, \qquad r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t, \tag{4.39}$$

for $t = n, \dots, 1$, with $r_n = 0$; this provides an efficient algorithm for calculating $\hat{\alpha}_1, \dots, \hat{\alpha}_n$. This smoother, together with the recursion for computing the variance matrix of the smoothed state vector which we present in Subsection 4.4.3, was proposed in the forms (4.39) and (4.43) below by de Jong (1988a), de Jong (1989) and Kohn and Ansley (1989) although the earlier treatments in the engineering literature by Bryson and Ho (1969) and Young (1984) are similar.

4.4.3 Smoothed state variance matrix

A recursion for calculating $V_t = \mathrm{Var}(\alpha_t|Y_n)$ will now be derived. We have defined $v_{t:n} = (v_t', \dots, v_n')'$. Applying Lemma 1 of Section 4.2 to the conditional joint distribution of α_t and $v_{t:n}$ given Y_{t-1}, taking $x = \alpha_t$ and $y = v_{t:n}$, we obtain from (4.3)

$$V_t = \mathrm{Var}(\alpha_t|Y_{t-1}, v_{t:n}) = P_t - \sum_{j=t}^{n} \mathrm{Cov}(\alpha_t, v_j) F_j^{-1} \mathrm{Cov}(\alpha_t, v_j)',$$

where $\mathrm{Cov}(\alpha_t, vj)$ and F_j are as in (4.31), since $v_t, \dots, v_n$ are independent of each other and of Y_{t-1} with zero means. Using (4.32) and (4.33) we obtain immediately

$$\begin{aligned} V_t &= P_t - P_t Z_t' F_t^{-1} Z_t P_t - P_t L_t' Z_{t+1}' F_{t+1}^{-1} Z_{t+1} L_t P_t - \cdots \\ &\quad - P_t L_t' \cdots L_{n-1}' Z_n' F_n^{-1} Z_n L_{n-1} \cdots L_t P_t \\ &= P_t - P_t N_{t-1} P_t, \end{aligned}$$

where

$$\begin{aligned} N_{t-1} &= Z_t' F_t^{-1} Z_t + L_t' Z_{t+1}' F_{t+1}^{-1} Z_{t+1} L_t + \cdots \\ &\quad + L_t' \cdots L_{n-1}' Z_n' F_n^{-1} Z_n L_{n-1} \cdots L_t. \end{aligned} \tag{4.40}$$

We note that here, as in the previous subsection, we interpret $L_t' \cdots L_{n-1}'$ as I_m when $t = n$ and as L_{n-1}' when $t = n - 1$. The value at time t is

$$\begin{aligned} N_t &= Z_{t+1}' F_{t+1}^{-1} Z_{t+1} + L_{t+1}' Z_{t+2}' F_{t+2}^{-1} Z_{t+2} L_{t+1} + \cdots \\ &\quad + L_{t+1}' \cdots L_{n-1}' Z_n' F_n^{-1} Z_n L_{n-1} \cdots L_{t+1}. \end{aligned} \tag{4.41}$$

Substituting (4.41) into (4.40) we obtain the backwards recursion

$$N_{t-1} = Z_t' F_t^{-1} Z_t + L_t' N_t L_t, \qquad t = n, \dots, 1. \tag{4.42}$$

Noting from (4.41) that $N_{n-1} = Z_n' F_n^{-1} Z_n$ we deduce that recursion (4.42) is initialised with $N_n = 0$. Collecting these results, we find that V_t can be efficiently calculated by the recursion

$$N_{t-1} = Z_t' F_t^{-1} Z_t + L_t' N_t L_t, \qquad V_t = P_t - P_t N_{t-1} P_t, \tag{4.43}$$

for $t = n, \dots, 1$ with $N_n = 0$. Since $v_{t+1}, \dots, v_n$ are independent it follows from (4.37) and (4.41) that $N_t = \mathrm{Var}(r_t)$.

4.4.4 State smoothing recursion

For convenience we collect together the smoothing equations for the state vector,

$$\begin{aligned} r_{t-1} &= Z_t' F_t^{-1} v_t + L_t' r_t, & N_{t-1} &= Z_t' F_t^{-1} Z_t + L_t' N_t L_t, \\ \hat{\alpha}_t &= a_t + P_t r_{t-1}, & V_t &= P_t - P_t N_{t-1} P_t, \end{aligned} \qquad (4.44)$$

for $t = n, \dots, 1$ initialised with $r_n = 0$ and $N_n = 0$. We refer to these collectively as the *state smoothing recursion*. As noted earlier, Lemmas 2 to 4 imply that the recursion (4.44) is also valid for non-normal cases in the MVLUE sense and for Bayesian analyses. Taken together, the recursions (4.24) and (4.44) will be referred to as the *Kalman filter and smoother*. We see that the way the filtering and smoothing is performed is that we proceed forwards through the series using (4.24) and backwards through the series using (4.44) to obtain $\hat{\alpha}_t$ and V_t for $t = 1, \dots, n$. During the forward pass we need to store the quantities v_t, F_t, K_t, a_t and P_t for $t = 1, \dots, n$. Alternatively we can store a_t and P_t only and recalculate v_t, F_t and K_t using a_t and P_t but this is usually not done since the dimensions of v_t, F_t and K_t are usually small relative to a_t and P_t, so the extra storage required is small. In Table 4.3 we present the dimensions of the vectors and matrices of the smoothing equations of this section and Subsection 4.5.3.

4.4.5 Updating smoothed estimates

In many situations observations come in one at a time and we wish to update the smoothed estimates each time a new observation comes in. We shall develop recursions for doing this which are computationally more efficient than applying (4.44) repeatedly.

Denote the new observation by y_{n+1} and suppose that we wish to calculate $\hat{\alpha}_{t|n+1} = \mathrm{E}(\alpha_t|Y_{n+1})$. For convenience, we relabel $\hat{\alpha}_t = \mathrm{E}(\alpha_t|Y_n)$ as $\hat{\alpha}_{t|n}$. We have

$$\begin{aligned} \hat{\alpha}_{t|n+1} &= \mathrm{E}(\alpha_t|Y_n, v_{n+1}) \\ &= \hat{\alpha}_{t|n} + \mathrm{Cov}(\alpha_t, v_{n+1}) F_{n+1}^{-1} v_{n+1}, \end{aligned}$$

Table 4.3 Dimensions of smoothing recursions of Subsections 4.4.4 and 4.5.3.

Vector		Matrix	
r_t	$m \times 1$	N_t	$m \times m$
$\hat{\alpha}_t$	$m \times 1$	V_t	$m \times m$
u_t	$p \times 1$	D_t	$p \times p$
$\hat{\varepsilon}_t$	$p \times 1$		
$\hat{\eta}_t$	$r \times 1$		

by Lemma 1. From (4.32) and (4.33) we have

$$\mathrm{Cov}(\alpha_t, v_{n+1}) = P_t L_t' \dots L_n' Z_{n+1}',$$

giving

$$\hat{\alpha}_{t|n+1} = \hat{\alpha}_{t|n} + P_t L_t' \dots L_n' Z_{n+1}' F_{n+1}^{-1} v_{n+1}, \tag{4.45}$$

for $t = 1, \dots, n$. In addition, from (4.17),

$$\hat{\alpha}_{n+1|n+1} = a_{n+1} + P_{n+1} Z_{n+1}' F_{n+1}^{-1} v_{n+1}. \tag{4.46}$$

Now consider the updating of the smoothed state variance matrix $V_t = \mathrm{Var}(\alpha_t | Y_n)$. For convenience we relabel this as $V_{t|n}$. Let $V_{t|n+1} = \mathrm{Var}(\alpha_t | Y_{n+1})$. By Lemma 1,

$$\begin{aligned} V_{t|n+1} &= \mathrm{Var}(\alpha_t | Y_n, v_{n+1}) \\ &= \mathrm{Var}(\alpha_t | Y_n) - \mathrm{Cov}(\alpha_t, v_{n+1}) F_{n+1}^{-1} \mathrm{Cov}(\alpha_t, v_{n+1})' \\ &= V_{t|n} - P_t L_t' \dots L_n' Z_{n+1}' F_{n+1}^{-1} Z_{n+1} L_n \dots L_t P_t, \end{aligned} \tag{4.47}$$

for $t = 1, \dots, n$ with

$$V_{n+1|n+1} = P_{n+1} - P_{n+1} Z_{n+1}' F_{n+1}^{-1} Z_{n+1} P_{n+1}. \tag{4.48}$$

Let $b_{t|n+1} = L_t' \dots L_n'$ with $b_{n+1|n+1} = I_m$. Then $b_{t|n+1} = L_t' b_{t+1|n+1}$ for $t = n, \dots, 1$ and we can write recursions (4.45) and (4.47) in the compact forms

$$\hat{\alpha}_{t|n+1} = \hat{\alpha}_{t|n} + P_t b_{t|n+1} Z_{n+1}' F_{n+1}^{-1} v_{n+1}, \tag{4.49}$$

$$V_{t|n+1} = V_{t|n} - P_t b_{t|n+1} Z_{n+1}' F_{n+1}^{-1} Z_{n+1} b_{t|n+1}' P_t, \tag{4.50}$$

for $n = t, t+1, \dots$ with $\hat{\alpha}_{n|n} = a_n + P_n Z_n' F_n^{-1} v_n$ and $V_{n|n} = P_n - P_n Z_n' F_n^{-1} Z_n P_n$. Note that P_t, L_t, F_{n+1} and v_{n+1} are all readily available from the Kalman filter.

4.4.6 Fixed-point and fixed-lag smoothers

The fixed-point smoother $\hat{\alpha}_{t|n} = \mathrm{E}(\alpha_t | Y_n)$ for t fixed and $n = t+1, t+2, \dots$ is given directly by the recursion (4.49) and its error variance matrix is given by (4.50).

From (4.39) the fixed-lag smoother $\hat{\alpha}_{n-j|n} = \mathrm{E}(\alpha_{n-j} | Y_n)$ for j fixed at possible values $0, 1, \dots, n-1$ and $n = j+1, j+2, \dots$ is given by

$$\hat{\alpha}_{n-j|n} = a_{n-j} + P_{n-j} r_{n-j-1}, \tag{4.51}$$

where r_{n-j-1} is obtained from the backward recursion

$$r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t, \qquad t = n, \dots, n-j, \tag{4.52}$$

with $r_n = 0$. From (4.43) its error variance matrix is given by

$$V_{n-j|n} = P_{n-j} - P_{n-j} N_{n-j-1} P_{n-j}, \tag{4.53}$$

where N_{n-j-1} is obtained from the backward recursion

$$N_{t-1} = Z_t' F_t^{-1} Z_t + L_t' N_t L_t, \qquad t = n, \dots, n-j, \tag{4.54}$$

with $N_n = 0$, for $j = 0, 1, \dots, n-1$.

4.5 Disturbance smoothing

In this section we will derive recursions for computing the smoothed estimates $\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t | Y_n)$ and $\hat{\eta}_t = \mathrm{E}(\eta_t | Y_n)$ of the disturbance vectors ε_t and η_t given all the observations $y_1, \dots, y_n$. These estimates have a variety of uses, particularly for parameter estimation and diagnostic checking, as will be indicated in Sections 7.3 and 7.5.

4.5.1 Smoothed disturbances

Let $\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t | Y_n)$. By Lemma 1 we have

$$\hat{\varepsilon}_t = \mathrm{E}(\varepsilon_t | Y_{t-1}, v_t, \dots, v_n) = \sum_{j=t}^{n} \mathrm{E}(\varepsilon_t v_j') F_j^{-1} v_j, \qquad t = 1, \dots, n, \tag{4.55}$$

since $\mathrm{E}(\varepsilon_t | Y_{t-1}) = 0$ and ε_t and v_t are jointly independent of Y_{t-1}. It follows from (4.30) that $\mathrm{E}(\varepsilon_t v_j') = \mathrm{E}(\varepsilon_t x_j') Z_j' + \mathrm{E}(\varepsilon_t \varepsilon_j')$ with $\mathrm{E}(\varepsilon_t x_t') = 0$ for $t = 1, \dots, n$ and $j = t, \dots, n$. Therefore

$$\mathrm{E}(\varepsilon_t v_j') = \begin{cases} H_t, & j = t, \\ \mathrm{E}(\varepsilon_t x_j') Z_j', & j = t+1, \dots, n, \end{cases} \tag{4.56}$$

with

$$\begin{aligned} \mathrm{E}(\varepsilon_t x_{t+1}') &= -H_t K_t', \\ \mathrm{E}(\varepsilon_t x_{t+2}') &= -H_t K_t' L_{t+1}', \\ &\vdots \\ \mathrm{E}(\varepsilon_t x_n') &= -H_t K_t' L_{t+1}' \cdots L_{n-1}', \end{aligned} \tag{4.57}$$

which follow from (4.30), for $t = 1, \dots, n-1$. Note that here as elsewhere we interpret $L_{t+1}' \cdots L_{n-1}'$ as I_m when $t = n-1$ and as L_{n-1}' when $t = n-2$. Substituting (4.56) into (4.55) leads to

$$\begin{aligned}\hat{\varepsilon}_t &= H_t(F_t^{-1}v_t - K_t'Z_{t+1}'F_{t+1}^{-1}v_{t+1} - K_t'L_{t+1}'Z_{t+2}'F_{t+2}^{-1}v_{t+2} - \cdots \\ &\quad - K_t'L_{t+1}'\cdots L_{n-1}'Z_n'F_n^{-1}v_n) \\ &= H_t(F_t^{-1}v_t - K_t'r_t) \\ &= H_tu_t, \qquad t = n,\ldots,1, \end{aligned} \tag{4.58}$$

where r_t is defined in (4.37) and

$$u_t = F_t^{-1}v_t - K_t'r_t. \tag{4.59}$$

We refer to the vector u_t as the *smoothing error*.

The smoothed estimate of η_t is denoted by $\hat{\eta}_t = \mathrm{E}(\eta_t|Y_n)$ and analogously to (4.55) we have

$$\hat{\eta}_t = \sum_{j=t}^{n} \mathrm{E}(\eta_t v_j')F_j^{-1}v_j, \qquad t = 1,\ldots,n. \tag{4.60}$$

The relations (4.30) imply that

$$\mathrm{E}(\eta_t v_j') = \begin{cases} Q_tR_t'Z_{t+1}', & j = t+1, \\ \mathrm{E}(\eta_t x_j')Z_j', & j = t+2,\ldots,n, \end{cases} \tag{4.61}$$

with

$$\begin{aligned} \mathrm{E}(\eta_t x_{t+2}') &= Q_tR_t'L_{t+1}', \\ \mathrm{E}(\eta_t x_{t+3}') &= Q_tR_t'L_{t+1}'L_{t+2}', \\ &\vdots \\ \mathrm{E}(\eta_t x_n') &= Q_tR_t'L_{t+1}'\cdots L_{n-1}', \end{aligned} \tag{4.62}$$

for $t = 1,\ldots,n-1$. Substituting (4.61) into (4.60) and noting that $\mathrm{E}(\eta_t v_t') = 0$ leads to

$$\begin{aligned}\hat{\eta}_t &= Q_tR_t'\left(Z_{t+1}'F_{t+1}^{-1}v_{t+1} + L_{t+1}'Z_{t+2}'F_{t+2}^{-1}v_{t+2} + \cdots + L_{t+1}'\cdots L_{n-1}'Z_n'F_n^{-1}v_n\right) \\ &= Q_tR_t'r_t, \qquad t = n,\ldots,1, \end{aligned} \tag{4.63}$$

where r_t is obtained from (4.38). This result is useful as we will show in the next section but it also gives the vector r_t the interpretation as the 'scaled' smoothed estimator of η_t. Note that in many practical cases the matrix Q_tR_t' is diagonal or sparse. Equations (4.58) and (4.65) below were first given by de Jong (1988a) and Kohn and Ansley (1989). Equations (4.63) and (4.68) below were given by Koopman (1993). Earlier developments on disturbance smoothing have been reported by Kailath and Frost (1968).

4.5.2 Smoothed disturbance variance matrices

The error variance matrices of the smoothed disturbances are developed by the same approach that we used in Subsection 4.4.3 to derive the smoothed state variance matrix. Using Lemma 1 we have

$$\begin{aligned}
\operatorname{Var}(\varepsilon_t|Y_n) &= \operatorname{Var}(\varepsilon_t|Y_{t-1}, v_t, \ldots, v_n) \\
&= \operatorname{Var}(\varepsilon_t|Y_{t-1}) - \sum_{j=t}^{n} \operatorname{Cov}(\varepsilon_t, v_j) \operatorname{Var}(v_j)^{-1} \operatorname{Cov}(\varepsilon_t, v_j)' \\
&= H_t - \sum_{j=t}^{n} \operatorname{Cov}(\varepsilon_t, v_j) F_j^{-1} \operatorname{Cov}(\varepsilon_t, v_j)', \qquad (4.64)
\end{aligned}$$

where $\operatorname{Cov}(\varepsilon_t, v_j) = \operatorname{E}(\varepsilon_t v_j')$ which is given by (4.56). By substitution we obtain

$$\begin{aligned}
\operatorname{Var}(\varepsilon_t|Y_n) &= H_t - H_t \big(F_t^{-1} + K_t' Z_{t+1}' F_{t+1}^{-1} Z_{t+1} K_t \\
&\qquad + K_t' L_{t+1}' Z_{t+2}' F_{t+2}^{-1} Z_{t+2} L_{t+1} K_t + \cdots \\
&\qquad + K_t' L_{t+1}' \cdots L_{n-1}' Z_n' F_n^{-1} Z_n L_{n-1} \cdots L_{t+1} K_t \big) H_t' \\
&= H_t - H_t \left(F_t^{-1} + K_t' N_t K_t \right) H_t \\
&= H_t - H_t D_t H_t, \qquad (4.65)
\end{aligned}$$

with

$$D_t = F_t^{-1} + K_t' N_t K_t, \qquad (4.66)$$

where N_t is defined in (4.41) and can be obtained from the backward recursion (4.42).

In a similar way the variance matrix $\operatorname{Var}(\eta_t|Y_n)$ is given by

$$\operatorname{Var}(\eta_t|Y_n) = \operatorname{Var}(\eta_t) - \sum_{j=t}^{n} \operatorname{Cov}(\eta_t, v_j) F_j^{-1} \operatorname{Cov}(\eta_t, v_j)', \qquad (4.67)$$

where $\operatorname{Cov}(\eta_t, v_j) = \operatorname{E}(\eta_t, v_j')$ which is given by (4.61). Substitution gives

$$\begin{aligned}
\operatorname{Var}(\eta_t|Y_n) &= Q_t - Q_t R_t' \big(Z_{t+1}' F_{t+1}^{-1} Z_{t+1} + L_{t+1}' Z_{t+2}' F_{t+2}^{-1} Z_{t+2} L_{t+1} + \cdots \\
&\qquad + L_{t+1}' \cdots L_{n-1}' Z_n' F_n^{-1} Z_n L_{n-1} \cdots L_{t+1} \big) R_t Q_t \\
&= Q_t - Q_t R_t' N_t R_t Q_t, \qquad (4.68)
\end{aligned}$$

where N_t is obtained from (4.42).

4.5.3 Disturbance smoothing recursion

For convenience we collect together the smoothing equations for the disturbance vectors,

$$\begin{aligned}
\hat{\varepsilon}_t &= H_t\left(F_t^{-1}v_t - K_t'r_t\right), & \mathrm{Var}(\varepsilon_t|Y_n) &= H_t - H_t\left(F_t^{-1} + K_t'N_tK_t\right)H_t,\\
\hat{\eta}_t &= Q_tR_t'r_t, & \mathrm{Var}(\eta_t|Y_n) &= Q_t - Q_tR_t'N_tR_tQ_t,\\
r_{t-1} &= Z_t'F_t^{-1}v_t + L_t'r_t, & N_{t-1} &= Z_t'F_t^{-1}Z_t + L_t'N_tL_t,
\end{aligned} \tag{4.69}$$

for $t = n, \dots, 1$ where $r_n = 0$ and $N_n = 0$. These equations can be reformulated as

$$\begin{aligned}
\hat{\varepsilon}_t &= H_tu_t, & \mathrm{Var}(\varepsilon_t|Y_n) &= H_t - H_tD_tH_t,\\
\hat{\eta}_t &= Q_tR_t'r_t, & \mathrm{Var}(\eta_t|Y_n) &= Q_t - Q_tR_t'N_tR_tQ_t,\\
u_t &= F_t^{-1}v_t - K_t'r_t, & D_t &= F_t^{-1} + K_t'N_tK_t,\\
r_{t-1} &= Z_t'u_t + T_t'r_t, & N_{t-1} &= Z_t'D_tZ_t + T_t'N_tT_t - Z_t'K_t'N_tT_t - T_t'N_tK_tZ_t,
\end{aligned}$$

for $t = n, \dots, 1$, which are computationally more efficient since they rely directly on the system matrices Z_t and T_t which have the property that they usually contain many zeros and ones. We refer to these equations collectively as the *disturbance smoothing recursion.* The smoothing error u_t and vector r_t are important in their own right for a variety of reasons which we will discuss in Section 7.5. The dimensions of the vectors and matrices of disturbance smoothing are given in Table 4.3.

We have shown that disturbance smoothing is performed in a similar way to state smoothing: we proceed forward through the series using (4.24) and backward through the series using (4.69) to obtain $\hat{\varepsilon}_t$ and $\hat{\eta}_t$ together with the corresponding conditional variances for $t = 1, \dots, n$. The storage requirement for (4.69) during the forward pass is less than for the state smoothing recursion (4.44) since here we only need v_t, F_t and K_t of the Kalman filter. Also, the computations are quicker for disturbance smoothing since they do not involve the vector a_t and the matrix P_t which are not sparse.

4.6 Other state smoothing algorithms

4.6.1 Classical state smoothing

Alternative algorithms for state smoothing have also been proposed. For example, Anderson and Moore (1979) present the so-called *classical fixed-interval smoother*, due to Rauch, Tung and Striebel (1965), which for our state space model is given by

$$\hat{\alpha}_t = a_{t|t} + P_{t|t}T_t'P_{t+1}^{-1}(\hat{\alpha}_{t+1} - a_{t+1}), \qquad t = n, \dots, 1, \tag{4.70}$$

where

$$a_{t|t} = \mathrm{E}(\alpha_t|Y_t) = a_t + P_t Z_t' F_t^{-1} v_t, \qquad P_{t|t} = \mathrm{Var}(\alpha_t|Y_t) = P_t - P_t Z_t' F_t^{-1} Z_t P_t;$$

see equations (4.17) and (4.18). Notice that $T_t P_{t|t} = L_t P_t$.

Following Koopman (1998), we now show that (4.39) can be derived from (4.70). Substituting for $a_{t|t}$ and $T_t P_{t|t}$ into (4.70) we have

$$\hat{\alpha}_t = a_t + P_t Z_t' F_t^{-1} v_t + P_t L_t' P_{t+1}^{-1}(\hat{\alpha}_{t+1} - a_{t+1}).$$

By defining $r_t = P_{t+1}^{-1}(\hat{\alpha}_{t+1} - a_{t+1})$ and re-ordering the terms, we obtain

$$P_t^{-1}(\hat{\alpha}_t - a_t) = Z_t' F_t^{-1} v_t + L_t' P_{t+1}^{-1}(\hat{\alpha}_{t+1} - a_{t+1}),$$

and hence

$$r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t,$$

which is (4.38). Note that the alternative definition of r_t also implies that $r_n = 0$. Finally, it follows immediately from the definitional relation $r_{t-1} = P_t^{-1}(\hat{\alpha}_t - a_t)$ that $\hat{\alpha}_t = a_t + P_t r_{t-1}$.

A comparison of the two different algorithms shows that the Anderson and Moore smoother requires inversion of $n-1$ possibly large matrices P_t whereas the smoother (4.39) requires no inversion other than of F_t which has been inverted as part of the computations of the Kalman filter. This is a considerable advantage for large models. For both smoothers the Kalman filter vector a_t and matrix P_t need to be stored together with v_t, F_t^{-1} and K_t, for $t = 1, \ldots, n$. The state smoothing equation of Koopman (1993), which we consider in Subsection 4.6.2, does not involve a_t and P_t and it therefore leads to further computational savings.

4.6.2 Fast state smoothing

The smoothing recursion for the disturbance vector η_t in Subsection 4.5.3 is particularly useful since it leads to a computationally more efficient method of calculating $\hat{\alpha}_t$ for $t = 1, \ldots, n$ than (4.39). Given the state equation

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t,$$

it follows by taking expectations given Y_n that

$$\begin{aligned} \hat{\alpha}_{t+1} &= T_t \hat{\alpha}_t + R_t \hat{\eta}_t \\ &= T_t \hat{\alpha}_t + R_t Q_t R_t' r_t, \qquad t = 1, \ldots, n, \end{aligned} \tag{4.71}$$

which is initialised via the relation (4.35) for $t = 1$, that is, $\hat{\alpha}_1 = a_1 + P_1 r_0$ where r_0 is obtained from (4.38). This recursion, due to Koopman (1993), can be used to generate the smoothed states $\hat{\alpha}_1, \dots, \hat{\alpha}_n$ by an algorithm different from (4.39); it does not require the storage of a_t and P_t and it does not involve multiplications by the full matrix P_t, for $t = 1, \dots, n$. After the Kalman filter and the storage of v_t, F_t^{-1} and K_t has taken place, the backwards recursion (4.38) is undertaken and the vector r_t is stored for which the storage space of K_t can be used so no additional storage space is required. It should be kept in mind that the matrices T_t and $R_t Q_t R_t'$ are usually sparse matrices containing many zero and unity values which make the application of (4.71) rapid; this property does not apply to P_t which is a full variance matrix. This approach, however, cannot be used to obtain a recursion for the calculation of $V_t = \text{Var}(\alpha_t|Y_n)$; if V_t is required then (4.39) and (4.43) should be used.

4.6.3 The Whittle relation between smoothed estimates

Whittle (1991) has provided an interesting relationship between smoothed state vector estimates based on direct likelihood arguments. In particular, for the local level model (2.3) this relationship reduces to

$$\hat{\alpha}_{t-1} = 2\hat{\alpha}_t - \hat{\alpha}_{t+1} - q\left(y_t - \hat{\alpha}_t\right),$$

for $t = n, n-1, \dots, 1$ with $\hat{\alpha}_{n+1} = \hat{\alpha}_n$. The initialisation and recursive equations follow directly from the loglikelihood function written as a joint density for $\alpha_1, \dots, \alpha_n, \alpha_{n+1}$ and $y_1, \dots, y_n$. Taking first derivatives with respect to α_t and solving the equations when set to zero yields the result. The results also hold for the general model. The recursive algorithm is appealing since the storage of Kalman filter quantities is not needed; the computation of $\hat{\alpha}_{n+1} = a_{n+1|n}$ is only required by the Kalman filter. However the algorithm is numerically unstable due to the accumulation of numerical inaccuracies.

4.6.4 Two filter formula for smoothing

Mayne (1966), Fraser and Potter (1969) and Kitagawa (1994) have developed a method of smoothing based on reversing the observations and the state space model equations (4.12). Reversion implies that the observation sequence $y_1, \dots, y_n$ becomes

$$y_1^-, \dots, y_n^- \;\equiv\; y_n, \dots, y_1.$$

We define the vector $Y_t^- = \left(y_1^{-\prime}, \dots, y_t^{-\prime}\right)'$ for $t = 1, \dots, n$. Here, Y_1^- is the $p \times 1$ vector y_n while Y_t and Y_t^- are two vectors which have different number of elements in almost all cases. The vectors Y_n and Y_n^- consist of the same elements but the elements are ordered in opposite directions. It is suggested that the same state space model can be considered for the time series in reverse order, that is

$$y_t^- = Z_t^- \alpha_t^- + \varepsilon_t^-, \qquad \alpha_{t+1}^- = T_t^- \alpha_t^- + R_t^- \eta_t^-, \tag{4.72}$$

for $t = 1, \dots, n$ where x_t^- is the t-th last element of the sequence $x_1, \dots, x_n$ for any variable x. It seems hard to justify this approach generally. However, in the case that the initial state vector is fully diffuse, that is $\alpha_1 \sim \mathrm{N}(0, \kappa I)$ with $\kappa \to \infty$, it can be justified. We illustrate this for the local level model of Chapter 2 using its matrix representation.

When the $n \times 1$ time series vector Y_n is generated by the univariate local level model, we have from (2.4) that $Y_n \sim \mathrm{N}(a_1 \mathtt{1}, \Omega)$ where a_1 is the initial state and Ω is specified as in (2.4) and (2.5). Let J here be the $n \times n$ matrix with its $(i, n-i+1)$ element equal to unity for $i = 1, \dots, n$ and with other elements equal to zero. Vector JY_n is then equal to Y_n in reverse order, that is Y_n^-. Matrix J has the properties $J' = J^{-1} = J$ and $J\,J = I$. In Subsection 2.3.1 it is argued that the Kalman filter effectively carries out the computation $Y_n'\Omega^{-1}Y_n = v'F^{-1}v$ via the Cholesky decomposition. When this computation can also be done in reverse order, it is implied that

$$Y_n'\Omega^{-1}Y_n = (JY_n)'\,\Omega^{-1}JY_n = Y_n'J\Omega^{-1}JY_n.$$

Hence we have implied that $\Omega^{-1} = J\Omega^{-1}J$. This property does not hold for a symmetric matrix generally. However, if Ω^{-1} is a symmetric Toeplitz matrix, the property holds. When matrix Ω is defined as (2.4), that is $\Omega = \mathtt{1}\mathtt{1}'P_1 + \Sigma$ where P_1 is the initial state variance and Σ is defined in (2.5), and when $P_1 = \kappa \to \infty$, Ω^{-1} converges to a symmetric Toeplitz matrix and hence the property holds. We can therefore reverse the ordering of the local level observations and obtain the same value for $v'F^{-1}v$. This argument can be generalised for state space models where the state vector is fully diffuse. In other cases, the justification of reversing the observations is harder to establish.

When reversing the observations is valid, the following two filtering methods for smoothing can be applied. The smoothed density of the state vector can be expressed by

$$\begin{aligned} p(\alpha_t|Y_n) &= p(\alpha_t|Y_{t-1}, Y_t^-) \\ &= c\,p(\alpha_t, Y_t^-|Y_{t-1}) \\ &= c\,p(\alpha_t|Y_{t-1})\,p(Y_t^-|\alpha_t, Y_{t-1}) \\ &= c\,p(\alpha_t|Y_{t-1})\,p(Y_t^-|\alpha_t), \end{aligned} \tag{4.73}$$

where c is some constant that does not depend on α_t. The last equality (4.73) holds since the disturbances $\varepsilon_t, \dots, \varepsilon_n$ and $\eta_t, \dots, \eta_{n-1}$ of (4.12) do not depend on Y_{t-1}. Taking logs of (4.73), multiplying it by -2 and only writing the terms associated with α_t, gives the equality

$$(\alpha_t - a_{t|n})'P_{t|n}^{-1}(\alpha_t - a_{t|n}) = (\alpha_t - a_t)'P_t^{-1}(\alpha_t - a_t) + (\alpha_t - a_{t|t}^-)'Q_{t|t}^{-1}(\alpha_t - a_{t|t}^-),$$

where

$$a_{t|t}^- = \mathrm{E}(\alpha_t|Y_t^-) = \mathrm{E}(\alpha_t|y_t, \dots, y_n), \qquad Q_{t|t} = \mathrm{Var}(\alpha_t|Y_t^-) = \mathrm{Var}(\alpha_t|y_t, \dots, y_n).$$

Vector $a_{t|t}^-$, associated with $a_{t|t}$, and matrix $Q_{t|t}$, associated with $P_{t|t}$, are obtained from the Kalman filter applied to the observations in reverse order, that is, $y_1^-, \dots, y_n^-$, and based on the model (4.72). By some minor matrix manipulations of the above equality, we obtain

$$a_{t|n} = P_{t|n}(P_t^{-1}a_t + Q_{t|t}^{-1}a_{t|t}^-), \qquad P_{t|n} = \left(P_t^{-1} + Q_{t|t}^{-1}\right)^{-1}, \tag{4.74}$$

for $t = 1, \dots, n$. To avoid taking inverses of the variance matrices, the application of an information filter can be considered; see Anderson and Moore (1979, Chapter 3) for a discussion of the information filter.

4.7 Covariance matrices of smoothed estimators

In this section we develop expressions for the covariances between the errors of the smoothed estimators $\hat{\varepsilon}_t$, $\hat{\eta}_t$ and $\hat{\alpha}_t$ contemporanously and for all leads and lags.

It turns out that the covariances of smoothed estimators rely basically on the cross-expectations $\mathrm{E}(\varepsilon_t r_j')$, $\mathrm{E}(\eta_t r_j')$ and $\mathrm{E}(\alpha_t r_j')$ for $j = t+1, \dots, n$. To develop these expressions we collect from equations (4.56), (4.57), (4.61), (4.62), (4.33) and (4.32) the results

$$\begin{aligned}
\mathrm{E}(\varepsilon_t x_t') &= 0, & \mathrm{E}(\varepsilon_t v_t') &= H_t, \\
\mathrm{E}(\varepsilon_t x_j') &= -H_t K_t' L_{t+1}' \cdots L_{j-1}', & \mathrm{E}(\varepsilon_t v_j') &= \mathrm{E}(\varepsilon_t x_j') Z_j', \\
\mathrm{E}(\eta_t x_t') &= 0, & \mathrm{E}(\eta_t v_t') &= 0, \\
\mathrm{E}(\eta_t x_j') &= Q_t R_t' L_{t+1}' \cdots L_{j-1}', & \mathrm{E}(\eta_t v_j') &= \mathrm{E}(\eta_t x_j') Z_j', \\
\mathrm{E}(\alpha_t x_t') &= P_t, & \mathrm{E}(\alpha_t v_t') &= P_t Z_t', \\
\mathrm{E}(\alpha_t x_j') &= P_t L_t' L_{t+1}' \cdots L_{j-1}', & \mathrm{E}(\alpha_t v_j') &= \mathrm{E}(\alpha_t x_j') Z_j',
\end{aligned} \tag{4.75}$$

for $j = t+1, \dots, n$. For the case $j = t+1$, we replace $L_{t+1}' \cdots L_t'$ by the identity matrix I.

We derive the cross-expectations below using the definitions

$$\begin{aligned}
r_j &= \sum_{k=j+1}^{n} L_{j+1}' \cdots L_{k-1}' Z_k' F_k^{-1} v_k, \\
N_j &= \sum_{k=j+1}^{n} L_{j+1}' \cdots L_{k-1}' Z_k' F_k^{-1} Z_k L_{k-1} \cdots L_{j+1},
\end{aligned}$$

which are given by (4.36) and (4.40), respectively. It follows that

$$\begin{aligned}
\mathrm{E}(\varepsilon_t r_j') &= \mathrm{E}(\varepsilon_t v_{j+1}')F_{j+1}^{-1}Z_{j+1} + \mathrm{E}(\varepsilon_t v_{j+2}')F_{j+2}^{-1}Z_{j+2}L_{j+1} + \cdots \\
&\quad + \mathrm{E}(\varepsilon_t v_n')F_n^{-1}Z_n L_{n-1}\cdots L_{j+1} \\
&= -H_t K_t' L_{t+1}' \cdots L_j' Z_{j+1}' F_{j+1}^{-1} Z_{j+1} \\
&\quad - H_t K_t' L_{t+1}' \cdots L_{j+1}' Z_{j+2}' F_{j+2}^{-1} Z_{j+2} L_{j+1} - \cdots \\
&\quad - H_t K_t' L_{t+1}' \cdots L_{n-1}' Z_n' F_n^{-1} Z_n L_{n-1} \cdots L_{j+1} \\
&= -H_t K_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j,
\end{aligned} \tag{4.76}$$

$$\begin{aligned}
\mathrm{E}(\eta_t r_j') &= \mathrm{E}(\eta_t v_{j+1}')F_{j+1}^{-1}Z_{j+1} + \mathrm{E}(\eta_t v_{j+2}')F_{j+2}^{-1}Z_{j+2}L_{j+1} + \cdots \\
&\quad + \mathrm{E}(\eta_t v_n')F_n^{-1}Z_n L_{n-1}\cdots L_{j+1} \\
&= Q_t R_t' L_{t+1}' \cdots L_j' Z_{j+1}' F_{j+1}^{-1} Z_{j+1} \\
&\quad + Q_t R_t' L_{t+1}' \cdots L_{j+1}' Z_{j+2}' F_{j+2}^{-1} Z_{j+2} L_{j+1} + \cdots \\
&\quad + Q_t R_t' L_{t+1}' \cdots L_{n-1}' Z_n' F_n^{-1} Z_n L_{n-1} \cdots L_{j+1} \\
&= Q_t R_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j,
\end{aligned} \tag{4.77}$$

$$\begin{aligned}
\mathrm{E}(\alpha_t r_j') &= \mathrm{E}(\alpha_t v_{j+1}')F_{j+1}^{-1}Z_{j+1} + \mathrm{E}(\alpha_t v_{j+2}')F_{j+2}^{-1}Z_{j+2}L_{j+1} + \cdots \\
&\quad + \mathrm{E}(\alpha_t v_n')F_n^{-1}Z_n L_{n-1}\cdots L_{j+1} \\
&= P_t L_t' L_{t+1}' \cdots L_j' Z_{j+1}' F_{j+1}^{-1} Z_{j+1} \\
&\quad + P_t L_t' L_{t+1}' \cdots L_{j+1}' Z_{j+2}' F_{j+2}^{-1} Z_{j+2} L_{j+1} + \cdots \\
&\quad + P_t L_t' L_{t+1}' \cdots L_{n-1}' Z_n' F_n^{-1} Z_n L_{n-1} \cdots L_{j+1} \\
&= P_t L_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j,
\end{aligned} \tag{4.78}$$

for $j = t, \ldots, n$. Hence

$$\begin{aligned}
\mathrm{E}(\varepsilon_t r_j') &= \mathrm{E}(\varepsilon_t x_{t+1}')N_{t+1,j}^*, \\
\mathrm{E}(\eta_t r_j') &= \mathrm{E}(\eta_t x_{t+1}')N_{t+1,j}^*, \\
\mathrm{E}(\alpha_t r_j') &= \mathrm{E}(\alpha_t x_{t+1}')N_{t+1,j}^*,
\end{aligned} \tag{4.79}$$

where $N_{t,j}^* = L_t' \cdots L_{j-1}' L_j' N_j$ for $j = t, \ldots, n$.

The cross-expectations of ε_t, η_t and α_t between the smoothed estimators

$$\hat{\varepsilon}_j = H_j\big(F_j^{-1}v_j - K_j' r_j\big), \qquad \hat{\eta}_j = Q_j R_j' r_j, \qquad \alpha_j - \hat{\alpha}_j = x_j - P_j r_{j-1},$$

for $j = t+1, \ldots, n$, are given by

$$\begin{aligned}
\mathrm{E}(\varepsilon_t \hat{\varepsilon}_j') &= \mathrm{E}(\varepsilon_t v_j') F_j^{-1} H_j - \mathrm{E}(\varepsilon_t r_j') K_j H_j, \\
\mathrm{E}(\varepsilon_t \hat{\eta}_j') &= \mathrm{E}(\varepsilon_t r_j') R_j Q_j, \\
\mathrm{E}[\varepsilon_t (\alpha_j - \hat{\alpha}_j)'] &= \mathrm{E}(\varepsilon_t x_j') - \mathrm{E}(\varepsilon_t r_{j-1}') P_j, \\
\mathrm{E}(\eta_t \hat{\varepsilon}_j') &= \mathrm{E}(\eta_t v_j') F_j^{-1} H_j - \mathrm{E}(\eta_t r_j') K_j H_j, \\
\mathrm{E}(\eta_t \hat{\eta}_j') &= \mathrm{E}(\eta_t r_j') R_j Q_j, \\
\mathrm{E}[\eta_t (\alpha_j - \hat{\alpha}_j)'] &= \mathrm{E}(\eta_t x_j') - \mathrm{E}(\eta_t r_{j-1}') P_j, \\
\mathrm{E}(\alpha_t \hat{\varepsilon}_j') &= \mathrm{E}(\alpha_t v_j') F_j^{-1} H_j - \mathrm{E}(\alpha_t r_j') K_j H_j, \\
\mathrm{E}(\alpha_t \hat{\eta}_j') &= \mathrm{E}(\alpha_t r_j') R_j Q_j, \\
\mathrm{E}[\alpha_t (\alpha_j - \hat{\alpha}_j)'] &= \mathrm{E}(\alpha_t x_j') - \mathrm{E}(\alpha_t r_{j-1}') P_j,
\end{aligned}$$

into which the expressions in equations (4.75), (4.76), (4.77) and (4.78) can be substituted.

The covariance matrices of the smoothed estimators at different times are derived as follows. We first consider the covariance matrix for the smoothed disturbance vector $\hat{\varepsilon}_t$, that is, $\mathrm{Cov}(\varepsilon_t - \hat{\varepsilon}_t, \varepsilon_j - \hat{\varepsilon}_j)$ for $t = 1, \ldots, n$ and $j = t+1, \ldots, n$. Since

$$\mathrm{E}[\hat{\varepsilon}_t (\varepsilon_j - \hat{\varepsilon}_j)'] = \mathrm{E}[\mathrm{E}\{\hat{\varepsilon}_t (\varepsilon_j - \hat{\varepsilon}_j)' | Y_n\}] = 0,$$

we have

$$\begin{aligned}
\mathrm{Cov}(\varepsilon_t - \hat{\varepsilon}_t, \varepsilon_j - \hat{\varepsilon}_j) &= \mathrm{E}[\varepsilon_t (\varepsilon_j - \hat{\varepsilon}_j)'] \\
&= -\mathrm{E}(\varepsilon_t \hat{\varepsilon}_j') \\
&= H_t K_t' L_{t+1}' \cdots L_{j-1}' Z_j' F_j^{-1} H_j \\
&\quad + H_t K_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j K_j H_j \\
&= H_t K_t' L_{t+1}' \cdots L_{j-1}' W_j',
\end{aligned}$$

where

$$W_j = H_j \big(F_j^{-1} Z_j - K_j' N_j L_j \big), \tag{4.80}$$

for $j = t+1, \ldots, n$. In a similar way obtain

$$\begin{aligned}
\mathrm{Cov}(\eta_t - \hat{\eta}_t, \eta_j - \hat{\eta}_j) &= -\mathrm{E}(\eta_t \hat{\eta}_j') \\
&= -Q_t R_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j R_j Q_j,
\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(\alpha_t - \hat{\alpha}_t, \alpha_j - \hat{\alpha}_j) &= -\mathrm{E}[\alpha_t(\alpha_j - \hat{\alpha}_j)'] \\ &= P_t L_t' L_{t+1}' \cdots L_{j-1}' - P_t L_t' L_{t+1}' \cdots L_{j-1}' N_{j-1} P_j \\ &= P_t L_t' L_{t+1}' \cdots L_{j-1}'(I - N_{j-1} P_j),\end{aligned}$$

for $j = t+1, \ldots, n$.

The cross-covariance matrices of the smoothed disturbances are obtained as follows. We have

$$\begin{aligned}\mathrm{Cov}(\varepsilon_t - \hat{\varepsilon}_t, \eta_j - \hat{\eta}_j) &= \mathrm{E}[(\varepsilon_t - \hat{\varepsilon}_t)(\eta_j - \hat{\eta}_j)'] \\ &= \mathrm{E}[\varepsilon_t(\eta_j - \hat{\eta}_j)'] \\ &= -\mathrm{E}(\varepsilon_t \hat{\eta}_j') \\ &= H_t K_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j R_j Q_j,\end{aligned}$$

for $j = t, t+1, \ldots, n$, and

$$\begin{aligned}\mathrm{Cov}(\eta_t - \hat{\eta}_t, \varepsilon_j - \hat{\varepsilon}_j) &= -\mathrm{E}(\eta_t \hat{\varepsilon}_j') \\ &= -Q_t R_t' L_{t+1}' \cdots L_{j-1}' Z_j' F_j^{-1} H_j \\ &\quad + Q_t R_t' L_{t+1}' \cdots L_{j-1}' N_j' K_j H_j \\ &= -Q_t R_t' L_{t+1}' \cdots L_{j-1}' W_j',\end{aligned}$$

for $j = t+1, \ldots, n$. The matrix products $L_{t+1}' \ldots L_{j-1}' L_j'$ for $j = t$ and $L_{t+1}' \ldots L_{j-1}'$ for $j = t+1$ are assumed to be equal to the identity matrix.

The cross-covariances between the smoothed state vector and the smoothed disturbances are obtained in a similar way. We have

$$\begin{aligned}\mathrm{Cov}(\alpha_t - \hat{\alpha}_t, \varepsilon_j - \hat{\varepsilon}_j) &= -\mathrm{E}(\alpha_t \hat{\varepsilon}_j') \\ &= -P_t L_t' L_{t+1}' \cdots L_{j-1}' Z_j' F_j^{-1} H_j \\ &\quad + P_t L_t' L_{t+1}' \cdots L_{j-1}' N_j' K_j H_j \\ &= -P_t L_t' L_{t+1}' \cdots L_{j-1}' W_j', \\ \mathrm{Cov}(\alpha_t - \hat{\alpha}_t, \eta_j - \hat{\eta}_j) &= -\mathrm{E}(\alpha_t \hat{\eta}_j') \\ &= -P_t L_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j R_j Q_j,\end{aligned}$$

for $j = t, t+1, \ldots, n$, and

$$\begin{aligned}\mathrm{Cov}(\varepsilon_t - \hat{\varepsilon}_t, \alpha_j - \hat{\alpha}_j) &= \mathrm{E}[\varepsilon_t(\alpha_j - \hat{\alpha}_j)'] \\ &= -H_t K_t' L_{t+1}' \cdots L_{j-1}' \\ &\quad + H_t K_t' L_{t+1}' \cdots L_{j-1}' N_{j-1} P_j \\ &= -H_t K_t' L_{t+1}' \cdots L_{j-1}'(I - N_{j-1} P_j),\end{aligned}$$

Table 4.4 Covariances of smoothed estimators for $t = 1, \dots, n$.

$\hat{\varepsilon}_t$	$\hat{\varepsilon}_j$	$H_t K_t' L_{t+1}' \cdots L_{j-1}' W_j'$	$j > t$
	$\hat{\eta}_j$	$H_t K_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j R_j Q_j$	$j \geq t$
	$\hat{\alpha}_j$	$-H_t K_t' L_{t+1}' \cdots L_{j-1}' (I_m - N_{j-1} P_j)$	$j > t$
$\hat{\eta}_t$	$\hat{\varepsilon}_j$	$-Q_t R_t' L_{t+1}' \cdots L_{j-1}' W_j'$	$j > t$
	$\hat{\eta}_j$	$-Q_t R_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j R_j Q_j$	$j > t$
	$\hat{\alpha}_j$	$Q_t R_t' L_{t+1}' \cdots L_{j-1}' (I_m - N_{j-1} P_j)$	$j > t$
$\hat{\alpha}_t$	$\hat{\varepsilon}_j$	$-P_t L_t' L_{t+1}' \cdots L_{j-1}' W_j'$	$j \geq t$
	$\hat{\eta}_j$	$-P_t L_t' L_{t+1}' \cdots L_{j-1}' L_j' N_j R_j Q_j$	$j \geq t$
	$\hat{\alpha}_j$	$P_t L_t' L_{t+1}' \cdots L_{j-1}' (I_m - N_{j-1} P_j)$	$j \geq t$

$$
\begin{aligned}
\mathrm{Cov}(\eta_t - \hat{\eta}_t, \alpha_j - \hat{\alpha}_j) &= \mathrm{E}[\eta_t(\alpha_j - \hat{\alpha}_j)'] \\
&= Q_t R_t' L_{t+1}' \cdots L_{j-1}' (I - N_{j-1} P_j),
\end{aligned}
$$

for $j = t + 1, \dots, n$.

The results here have been developed by de Jong and MacKinnon (1988), who derived the covariances between smoothed state vector estimators, and by Koopman (1993) who derived the covariances between the smoothed disturbance vectors estimators. The results in this section have also been reviewed by de Jong (1998). The auto- and cross-covariance matrices are for convenience collected in Table 4.4.

4.8 Weight functions

4.8.1 Introduction

Up to this point we have developed recursions for the evaluation of the conditional mean vector and variance matrix of the state vector α_t given the observations $y_1, \dots, y_{t-1}$ (prediction), given the observations $y_1, \dots, y_t$ (filtering) and given the observations $y_1, \dots, y_n$ (smoothing). We have also developed recursions for the conditional mean vectors and variance matrices of the disturbance vectors ε_t and η_t given the observation $y_1, \dots, y_n$. It follows that these conditional means are weighted sums of past (filtering), of past and present (contemporaneous filtering) and of all (smoothing) observations. It is of interest to study these weights to gain a better understanding of the properties of the estimators as is argued in Koopman and Harvey (2003). For example, the weights for the smoothed estimator of a trend component around time $t = n/2$, that is, in the middle of the series, should be symmetric and centred around t with exponentially declining weights unless specific circumstances require a different pattern. Models which produce weight patterns for the trend components which differ from what is regarded as appropriate should be investigated. In effect, the weights can be regarded as what are known as kernel functions in the field of nonparametric regression; see, for example, Green and Silverman (1994).

In the case when the state vector contains regression coefficients, the associated weights for the smoothed state vector can be interpreted as leverage statistics as studied in Cook and Weisberg (1982) and Atkinson (1985) in the context of regression models. Such statistics for state space models have been developed with the emphasis on the smoothed signal estimator $Z_t\hat{\alpha}_t$ by, for example, Kohn and Ansley (1989), de Jong (1989), Harrison and West (1991) and de Jong (1998). Since the concept of leverage is more useful in a regression context, we will refer to the expressions below as weights. Given the results of this chapter so far, it is straightforward to develop the weight expressions.

4.8.2 Filtering weights

It follows from the linear properties of the normal distribution that the filtered estimator of the state vector can be expressed as a weighted vector sum of past observations, that is

$$a_t = \sum_{j=1}^{t-1} \omega_{jt} y_j,$$

where ω_{jt} is an $m \times p$ matrix of weights associated with the estimator a_t and the jth observation. An expression for the weight matrix can be obtained using the fact that

$$\mathrm{E}(a_t \varepsilon_j') = \omega_{jt}\mathrm{E}(y_j \varepsilon_j') = \omega_{jt} H_j.$$

Since $x_t = \alpha_t - a_t$ and $\mathrm{E}(\alpha_t \varepsilon_j') = 0$, we can use (4.29) to obtain

$$\begin{aligned}\mathrm{E}(a_t \varepsilon_j') &= \mathrm{E}(x_t \varepsilon_j') = L_{t-1}\,\mathrm{E}(x_{t-1}\varepsilon_j') \\ &= L_{t-1}L_{t-2}\cdots L_{j+1}K_j H_j,\end{aligned}$$

which gives

$$\omega_{jt} = L_{t-1}L_{t-2}\cdots L_{j+1}K_j,$$

for $j = t-1, \ldots, 1$. In a similar way we can obtain the weights associated with other filtering estimators. In Table 4.5 we give a selection of such expressions from which the weights are obtained by disregarding the last matrix H_j. Finally, the expression for weights of $Z_t a_{t|t}$ follows since $Z_t P_t Z_t' = F_t - H_t$ and

$$Z_t\left(I - P_t Z_t' F_t^{-1} Z_t\right) = \left[I - (F_t - H_t)F_t^{-1}\right] Z_t = H_t F_t^{-1} Z_t.$$

Table 4.5 Expressions for $\mathrm{E}(s_t\varepsilon_j')$ with $1 \le t \le n$ given (filtering).

s_t	$j < t$	$j = t$	$j > t$
a_t	$L_{t-1}\cdots L_{j+1}K_jH_j$	0	0
$a_{t\vert t}$	$(I - P_tZ_t'F_t^{-1}Z_t)L_{t-1}\cdots L_{j+1}K_jH_j$	$P_tZ_t'F_t^{-1}H_t$	0
Z_ta_t	$Z_tL_{t-1}\cdots L_{j+1}K_jH_j$	0	0
$Z_ta_{t\vert t}$	$H_tF_t^{-1}Z_tL_{t-1}\cdots L_{j+1}K_jH_j$	$\left(I - H_tF_t^{-1}\right)H_t$	0
v_t	$-Z_tL_{t-1}\cdots L_{j+1}K_jH_j$	H_t	0

Table 4.6 Expressions for $\mathrm{E}(s_t\varepsilon_j')$ with $1 \le t \le n$ given (smoothing).

s_t	$j < t$	$j = t$	$j > t$
$\hat\varepsilon_t$	$-W_tL_{t-1}\cdots L_{j+1}K_jH_j$	$H_tD_tH_t$	$-H_tK_t'L_{t+1}'\cdots L_{j-1}'W_j'$
$\hat\eta_t$	$-Q_tR_t'N_tL_tL_{t-1}\cdots L_{j+1}K_jH_j$	$Q_tR_t'N_tK_tH_t$	$Q_tR_t'L_{t+1}'\cdots L_{j-1}'W_j'$
$\hat\alpha_t$	$(I-P_tN_{t-1})L_{t-1}\cdots L_{j+1}K_jH_j$	P_tW_t'	$P_tL_t'L_{t+1}'\cdots L_{j-1}'W_j'$
$Z_t\hat\alpha_t$	$W_tL_{t-1}\cdots L_{j+1}K_jH_j$	$(I-H_tD_t)H_t$	$H_tK_t'L_{t+1}'\cdots L_{j-1}'W_j'$

4.8.3 Smoothing weights

The weighting expressions for smoothing estimators can be obtained in a similar way to those used for filtering. For example, the smoothed estimator of the measurement disturbance vector can be expressed as a weighted vector sum of past, current and future observations, that is

$$\hat\varepsilon_t = \sum_{j=1}^{n} \omega_{jt}^{\varepsilon} y_j,$$

where $\omega_{jt}^{\varepsilon}$ is a $p \times p$ matrix of weights associated with the estimator $\hat\varepsilon_t$ and the jth observation. An expression for the weight matrix can be obtained using the fact that

$$\mathrm{E}(\hat\varepsilon_t\varepsilon_j') = \omega_{jt}^{\varepsilon}\,\mathrm{E}(y_j\varepsilon_j') = \omega_{jt}^{\varepsilon}H_j.$$

Expressions for the covariance matrices for smoothed disturbances are developed in Section 4.7 and they are directly related to the expression for $\mathrm{E}(\hat\varepsilon_t\varepsilon_j')$ because

$$\mathrm{Cov}(\varepsilon_t - \hat\varepsilon_t, \varepsilon_j - \hat\varepsilon_j) = \mathrm{E}[(\varepsilon_t - \hat\varepsilon_t)\varepsilon_j'] = -\mathrm{E}(\hat\varepsilon_t\varepsilon_j'),$$

with $1 \le t \le n$ and $j = 1, \ldots, n$. Therefore, no new derivations need to be given here and we only state the results as presented in Table 4.6.

For example, to obtain the weights for the smoothed estimator of α_t, we require

$$\begin{aligned}\mathrm{E}(\hat\alpha_t\varepsilon_j') &= -\mathrm{E}[(\alpha_t - \hat\alpha_t)\varepsilon_j'] = -\mathrm{E}[\varepsilon_j(\alpha_t - \hat\alpha_t)']' \\ &= \mathrm{Cov}(\varepsilon_j - \hat\varepsilon_j, \alpha_t - \hat\alpha_t)',\end{aligned}$$

for $j < t$. An expression for this latter quantity can be directly obtained from Table 4.4 but notice that the indices j and t of Table 4.4 need to be reversed here. Further,

$$\mathrm{E}(\hat\alpha_t\varepsilon_j') = \mathrm{Cov}(\alpha_t - \hat\alpha_t, \varepsilon_j - \hat\varepsilon_j)$$

for $j \ge t$ can also be obtained from Table 4.4. In the same way we can obtain the weights for the smoothed estimators of ε_t and η_t from Table 4.4 as reported in Table 4.6. Finally, the expression for weights of $Z_t\hat\alpha_t$ follows since

$Z_tP_tZ_t' = F_t - H_t$ and $Z_tP_tL_t' = H_tK_t'$. Hence, by using the equations (4.42), (4.66) and (4.80), we have

$$\begin{aligned} Z_t(I - P_tN_{t-1}) &= Z_t - Z_tP_tZ_t'F_t^{-1}Z_t + Z_tP_tL_t'N_tL_t \\ &= H_tF_t^{-1}Z_t + H_tK_t'N_tL_t = W_t, \\ Z_tP_tW_t' &= (Z_tP_tZ_t'F_t^{-1} - Z_tP_tL_t'N_tK_t)H_t \\ &= [(F_t - H_t)F_t^{-1} - H_tK_t'N_tK_t]H_t \\ &= (I - H_tD_t)H_t. \end{aligned}$$

4.9 Simulation smoothing

The drawing of samples of state or disturbance vectors conditional on the observations held fixed is called *simulation smoothing.* Such samples are useful for investigating the performance of techniques of analysis proposed for the linear Gaussian model, and for Bayesian analysis based on this model. The primary purpose of simulation smoothing in this book, however, will be to serve as the basis for the simulation techniques we shall develop in Part II for dealing with non-Gaussian and nonlinear models from both classical and Bayesian perspectives.

In this section we will show how to draw random samples of the disturbance vectors ε_t and η_t, and the state vector α_t, for $t = 1, \ldots, n$, generated by the linear Gaussian model (4.12) conditional on the observed vector y_n. The resulting algorithm is sometimes called a *forwards filtering, backwards sampling* algorithm.

Frühwirth-Schnatter (1994) and Carter and Kohn (1994) independently developed methods for simulation smoothing of the state vector based on the identity

$$p(\alpha_1, \ldots, \alpha_n|Y_n) = p(\alpha_n|Y_n)p(\alpha_{n-1}|Y_n, \alpha_n) \cdots p(\alpha_1|Y_n, \alpha_2, \ldots, \alpha_n). \quad (4.81)$$

de Jong and Shephard (1995) made significant progress by first concentrating on sampling the disturbances and subsequently sampling the states.

Subsequently, in Durbin and Koopman (2002), we developed a method which is based only on mean corrections of unconditional vectors and which is much simpler and computationally more efficient than the de Jong-Shephard and earlier procedures. The treatment which follows is based on this approach; the de Jong-Shephard method is summarised in Subsection 4.9.3.

4.9.1 Simulation smoothing by mean corrections

Our aim is to draw samples of the disturbances $\varepsilon_1, \ldots, \varepsilon_n$ and $\eta_1, \ldots, \eta_n$ given the observational set Y_n. Let $w = (\varepsilon_1', \eta_1', \ldots, \varepsilon_n', \eta_n')'$ and let $\widehat{w} = \mathrm{E}(w|Y_n)$, $W = \mathrm{Var}(w|Y_n)$. Since model (4.12) is linear and Gaussian, the conditional density of w given Y_n is $\mathrm{N}(\widehat{w}, W)$. The mean vector $\widehat{w}$ is easily calculated from recursions

(4.58) and (4.63); we show below that for the mean-correction method we do not need to calculate the variance matrix W, which is convenient computationally.

The unconditional distribution of w is

$$p(w) = \mathrm{N}(0, \Phi), \qquad \text{where} \quad \Phi = \mathrm{diag}(H_1, Q_1, \ldots, H_n, Q_n).$$

Let w^+ be a random vector drawn from $p(w)$. The process of drawing w^+ is straightforward, particularly since in many cases in practice the matrices H_t and Q_t for $t = 1, \ldots, n$ are scalar or diagonal. Denote by y^+ the stacked vector of values of y_t generated recursively by drawing a vector α_1^+ from $p(\alpha_1)$, assuming this density is known, and replacing α_1 and w in model (4.12) by α_1^+ and w^+. Compute $\widehat{w}^+ = \mathrm{E}(w|y^+)$ from recursions (4.58) and (4.63). It follows from Lemma 1 of Section 4.2 that the conditional variance matrix of a vector x given another vector y in a multivariate normal distribution does not depend on the value of y. Hence, $\mathrm{Var}(w|y^+) = W$ and conditionally on y^+ we have $w^+ - \widehat{w}^+ \sim \mathrm{N}(0, W)$. Since the density $\mathrm{N}(0, W)$ does not depend on y^+, it holds that $w^+ - \widehat{w}^+ \sim \mathrm{N}(0, W)$ unconditionally, and

$$\widetilde{w} = w^+ - \widehat{w}^+ + \widehat{w}, \tag{4.82}$$

is a random draw from $\mathrm{N}(\widehat{w}, W)$ as required. The simplicity of this expression for $\widetilde{w}$, and the elementary nature of the draw of w^+ from $\mathrm{N}(0, \Phi)$, together account for the greater efficiency of this approach to simulation smoothing compared with earlier methods. Of course, if draws of $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_n')'$ only or $\eta = (\eta_1', \ldots, \eta_n')'$ only are required we just replace w by ε or η as appropriate. It is interesting to note that the form of (4.82) is similar to expression (4) of Journel (1974); however, Journel's work was done in a different context. Also, we were unaware of it when developing the mean-corrections method in Durbin and Koopman (2002).

The above theory has been derived on the assumption that the initial vector α_1 has the distribution $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known. In practice, however, it is common for at least some of the elements of α_1 to be either fixed and unknown or to be random variables with arbitrarily large variances; such elements are called *diffuse*. The modifications to the theory that are required when some elements of α_1 are diffuse are given in Section 5.5.

4.9.2 Simulation smoothing for the state vector

To construct an algorithm for generating draws of the state vector $\alpha = (\alpha_1', \ldots, \alpha_n')'$ from the conditional density $p(\alpha|Y_n)$, we denote a draw from $p(\alpha)$ as α^+ and a draw from $p(\alpha|Y_n)$ as $\widetilde{\alpha}$. To generate α^+ we first draw w^+ as above and then use model (4.12) as a recursion initialised by $\alpha_1^+ \sim p(\alpha_1)$ with α and w replaced by α^+ and w^+, respectively, as in Subsection 4.9.1. We compute $\widehat{\alpha} = \mathrm{E}(\alpha|Y_n)$ and $\widehat{\alpha}^+ = \mathrm{E}(\alpha|y^+)$ by the Kalman filter and the smoothing recursions (4.58) and (4.63), finally using the forward recursion (4.71) to generate $\widehat{\alpha}$ and $\widehat{\alpha}^+$. The required draw of $\widetilde{\alpha}$ is then given by the expression $\widetilde{\alpha} = \alpha^+ - \widehat{\alpha}^+ + \widehat{\alpha}$.

4.9.3 de Jong–Shephard method for simulation of disturbances

The mean-corrections method for simulation smoothing works well in nearly all practical cases. However, there may be cases where the mean-corrections method cannot be implemented properly due to imposed ill-defined variance matrices; see the discussion in Jungbacker and Koopman (2007, §1). Since the de Jong–Shephard method does work generally, we present here the recursions developed in de Jong and Shephard (1995). We begin by presenting the recursions required for drawing a sample of the observation disturbances $\varepsilon_1, \dots, \varepsilon_n$ from the conditional density $p(\varepsilon_1, \dots, \varepsilon_n | Y_n)$. Let

$$\bar{\varepsilon}_t = \mathrm{E}(\varepsilon_t | \varepsilon_{t+1}, \dots, \varepsilon_n, Y_n), \qquad t = n-1, \dots, 1, \tag{4.83}$$

with $\bar{\varepsilon}_n = \mathrm{E}(\varepsilon_n | Y_n) = H_n F_n^{-1} v_n$. It can be shown that

$$\bar{\varepsilon}_t = H_t \left(F_t^{-1} v_t - K_t' \tilde{r}_t \right), \qquad t = n-1, \dots, 1, \tag{4.84}$$

where $\tilde{r}_t$ is determined by the backward recursion

$$\tilde{r}_{t-1} = Z_t' F_t^{-1} v_t - \tilde{W}_t' C_t^{-1} d_t + L_t' \tilde{r}_t, \qquad t = n, n-1, \dots, 1, \tag{4.85}$$

with $\tilde{r}_n = 0$ and

$$\tilde{W}_t = H_t \left(F_t^{-1} Z_t - K_t' \tilde{N}_t L_t \right), \tag{4.86}$$

$$\tilde{N}_{t-1} = Z_t' F_t^{-1} Z_t + \tilde{W}_t' C_t^{-1} \tilde{W}_t + L_t' \tilde{N}_t L_t, \tag{4.87}$$

for $t = n, n-1, \dots, 1$ and with $\tilde{N}_n = 0$. Here $C_t = \mathrm{Var}(\varepsilon_t - \bar{\varepsilon}_t)$ which is determined by

$$C_t = H_t - H_t \left(F_t^{-1} + K_t' \tilde{N}_t K_t \right) H_t, \qquad t = n, \dots, 1. \tag{4.88}$$

In these formulae, F_t, v_t and K_t are obtained from the Kalman filter (4.24) and $L_t = T_t - K_t Z_t$. The required draw from $p(\varepsilon_t | \varepsilon_{t+1}, \dots, \varepsilon_n, Y_n)$ is then obtained as a random draw from $\mathrm{N}(\bar{\varepsilon}_t, C_t)$.

We now present the recursions required for selecting a sample of state disturbances $\eta_1, \dots, \eta_n$ from density $p(\eta_1, \dots, \eta_n | Y_n)$. Let

$$\bar{\eta}_t = \mathrm{E}(\eta_t | \eta_{t+1}, \dots, \eta_n, Y_n), \qquad \bar{C}_t = \mathrm{Var}(\eta_t | Y_n, \eta_{t+1}, \dots, \eta_n),$$

for $t = n-1, \dots, 1$ with $\bar{\eta}_n = \mathrm{E}(\eta_n | Y_n) = 0$ and $\bar{C}_n = \mathrm{Var}(\eta_n | Y_n) = Q_n$. Further, let

$$\bar{W}_t = Q_t R_t' \tilde{N}_t L_t, \tag{4.89}$$

where $\tilde{N}_t$ is determined by the backward recursion (4.87) with $\tilde{W}_t$ replaced by $\bar{W}_t$ in (4.89). Then $\bar{\eta}_t$ is given by the relation

$$\bar{\eta}_t = Q_t R_t \tilde{r}_t, \qquad t = n-1, \dots, 1, \tag{4.90}$$

where $\tilde{r}_t$ is determined by the recursion (4.85) with $\tilde{W}_t$ replaced by $\bar{W}_t$. Furthermore, $\bar{C}_t$ is given by

$$\bar{C}_t = Q_t - Q_t R_t' \tilde{N}_t R_t Q_t, \qquad t = n, \dots, 1. \tag{4.91}$$

with $\tilde{N}_t$ as in (4.87). The required draw from $p(\eta_t|\eta_{t+1}, \dots, \eta_n, Y_n)$ is then obtained as a random draw from $\mathrm{N}(\bar{\eta}_t, \bar{C}_t)$ for $t = n-1, \dots, 1$ with $\eta_n \sim \mathrm{N}(0, Q_n)$.

By adopting the same arguments as we used for developing the quick state smoother in Subsection 4.6.2, we obtain the following forwards recursion for simulating from the conditional density $p(\alpha|Y_n)$ when the de Jong–Shephard method is used

$$\tilde{\alpha}_{t+1} = T_t \tilde{\alpha}_t + R_t \tilde{\eta}_t, \tag{4.92}$$

for $t = 1, \dots, n$ with $\tilde{\alpha}_1 = a_1 + P_1 \tilde{r}_0$.

The proofs of these recursions are long and intricate so they will not be given here. Instead, the reader is referred to the proofs in de Jong and Shephard (1995, §2) and Jungbacker and Koopman (2007, Theorem 2 and Proposition 6).

4.10 Missing observations

We now demonstrate that when the linear state space model (4.12) is used for the analysis, with or without the assumption of normality, allowance for missing observations in the derivation of the Kalman filter and smoother is particularly simple. Suppose that observations y_j are missing for $j = \tau, \dots, \tau^*$ with $1 < \tau < \tau^* < n$. An obvious procedure is to define a new series $y_{t^*}^*$ where $y_t^* = y_t$ for $t = t^* = 1, \dots, \tau - 1$ and $y_{t^*}^* = y_t$ for $t = \tau^* + 1, \dots, n$ and $t^* = \tau, \dots, n - (\tau^* - \tau)$. The model for $y_{t^*}^*$ is then the same as (4.12) where $y_t = y_{t^*}$, $\alpha_t = \alpha_{t^*}$ and the disturbances are associated with time index t^*. The system matrices remain associated with the time index t. The state update equation at time $t^* = \tau - 1$ is replaced by

$$\alpha_\tau = T^*_{\tau^*, \tau-1} \alpha_{\tau-1} + \eta^*_{\tau-1}, \qquad \eta^*_{\tau-1} \sim N\left(0, \sum_{j=\tau}^{\tau^*+1} T^*_{\tau^*,j} R_{j-1} Q_{j-1} R'_{j-1} T^{*\,\prime}_{\tau^*,j}\right),$$

with $T^*_{i,j} = T_i T_{i-1} \dots T_j$ for $j = \tau, \dots, \tau^*$ and $T^*_{\tau^*, \tau^*+1} = I_r$. Filtering and smoothing then proceed by the methods developed above for model (4.12). The procedure is extended in an obvious way when observations are missing at several points in the series.

It is, however, easier to proceed as follows. For $t = \tau, \dots, \tau^* - 1$ we have

$$\begin{aligned} a_{t|t} &= \mathrm{E}(\alpha_t|Y_t) = \mathrm{E}(\alpha_t|Y_{t-1}) \\ &= a_t, \\ P_{t|t} &= \mathrm{Var}(\alpha_t|Y_t) = \mathrm{Var}(\alpha_t|Y_{t-1}) \\ &= P_t, \\ a_{t+1} &= \mathrm{E}(\alpha_{t+1}|Y_t) = \mathrm{E}(T_t\alpha_t + R_t\eta_t|Y_{t-1}) \\ &= T_t a_t, \\ P_{t+1} &= \mathrm{Var}(\alpha_{t+1}|Y_t) = \mathrm{Var}(T_t\alpha_t + R_t\eta_t|Y_{t-1}) \\ &= T_t P_t T_t' + R_t Q_t R_t'. \end{aligned}$$

It follows that the Kalman filter for the case of missing observations is obtained simply by putting $Z_t = 0$ in (4.24) for $t = \tau, \dots, \tau^* - 1$; the same applies to (4.26) for the case of models with mean adjustments. Similarly, the backwards smoothing recursions (4.38) and (4.42) become

$$r_{t-1} = T_t' r_t, \qquad N_{t-1} = T_t' N_t T_t, \qquad t = \tau^* - 1, \dots, \tau; \qquad (4.93)$$

other relevant equations in (4.44) remain the same. It therefore follows that in smoothing as in filtering, we can use the same recursion (4.44) as when all observations are available by taking $Z_t = 0$ at time points where observations are missing. As with Kalman filtering and smoothing for complete sets of observations, the results remain valid for MLVUE and Bayesian analysis by Lemmas 2, 3 and 4. This simple treatment of missing observations is one of the attractions of the state space methods for time series analysis.

Suppose that at time t some but not all of the elements of the observation vector y_t are missing. Let y_t^* be the vector of values actually observed. Then $y_t^* = W_t y_t$ where W_t is a known matrix whose rows are a subset of the rows of I. Consequently, at time points where not all elements of y_t are available, the first equation of (4.12) is replaced by the equation

$$y_t^* = Z_t^* \alpha_t + \varepsilon_t^*, \qquad \varepsilon_t^* \sim \mathrm{N}(0, H_t^*),$$

where $Z_t^* = W_t Z_t$, $\varepsilon_t^* = W_t \varepsilon_t$ and $H_t^* = W_t H_t W_t'$. The Kalman filter and smoother then proceed exactly as in the standard case, provided that y_t, Z_t and H_t are replaced by y_t^*, Z_t^* and H_t^* at relevant time points. Of course, the dimensionality of the observation vector varies over time, but this does not affect the validity of the formulae; see also Section 4.12. The missing elements can be estimated by appropriate elements of $Z_t \hat{\alpha}_t$ where $\hat{\alpha}_t$ is the smoothed value. A more convenient method for dealing with missing elements for such multivariate models is given in Section 6.4 which is based on an element by element treatment of the observation vector y_t.

When observations or observational elements are missing, simulation samples obtained by the methods described in Section 4.9 carry through without further complexities. The mean correction methods are based on Kalman filtering and smoothing methods which can deal with missing values in the way shown in this subsection.

4.11 Forecasting

For many time series investigations, forecasting of future observations of the state vector is of special importance. In this section, we shall show that minimum mean square error forecasts can be obtained simply by treating future values of y_t as missing observations and using the techniques of the last section.

Suppose we have vector observations $y_1, \ldots, y_n$ which follow the state space model (4.12) and we wish to forecast y_{n+j} for $j = 1, \ldots, J$. For this purpose let us choose the estimate $\bar{y}_{n+j}$ which has a minimum mean square error matrix given Y_n, that is, $\bar{F}_{n+j} = \mathrm{E}[(\bar{y}_{n+j} - y_{n+j})(\bar{y}_{n+j} - y_{n+j})'|Y_n]$ is a minimum in the matrix sense for all estimates of y_{n+j}. It is standard knowledge that if x is a random vector with mean μ and finite variance matrix, then the value of a constant vector λ which minimises $\mathrm{E}[(\lambda - x)(\lambda - x)']$ is $\lambda = \mu$. It follows that the minimum mean square error forecast of y_{n+j} given Y_n is the conditional mean $\bar{y}_{n+j} = \mathrm{E}(Y_{n+j}|Y_n)$.

For $j = 1$ the forecast is straightforward. We have $y_{n+1} = Z_{n+1}\alpha_{n+1} + \varepsilon_{n+1}$ so

$$\begin{aligned}\bar{y}_{n+1} &= Z_{n+1}\,\mathrm{E}(\alpha_{n+1}|Y_n)\\ &= Z_{n+1}a_{n+1},\end{aligned}$$

where a_{n+1} is the estimate (4.21) of α_{n+1} produced by the Kalman filter. The conditional mean square error matrix

$$\begin{aligned}\bar{F}_{n+1} &= \mathrm{E}[(\bar{y}_{n+1} - y_{n+1})(\bar{y}_{n+1} - y_{n+1})'|Y_n]\\ &= Z_{n+1}P_{n+1}Z_{n+1}' + H_{n+1},\end{aligned}$$

is produced by the Kalman filter relation (4.16). We now demonstrate that we can generate the forecasts $\bar{y}_{n+j}$ for $j = 2, \ldots, J$ merely by treating $y_{n+1}, \ldots, y_{n+J}$ as missing values as in Section 4.10. Let $\bar{a}_{n+j} = \mathrm{E}(\alpha_{n+j}|Y_n)$ and $\bar{P}_{n+j} = \mathrm{E}[(\bar{a}_{n+j} - \alpha_{n+j})(\bar{a}_{n+j} - \alpha_{n+j})'|Y_n]$. Since $y_{n+j} = Z_{n+j}\alpha_{n+j} + \varepsilon_{n+j}$ we have

$$\begin{aligned}\bar{y}_{n+j} &= Z_{n+j}\,\mathrm{E}(\alpha_{n+j}|Y_n)\\ &= Z_{n+j}\bar{a}_{n+j},\end{aligned}$$

with conditional mean square error matrix

$$\begin{aligned}\bar{F}_{n+j} &= \mathrm{E}[\{Z_{n+j}(\bar{a}_{n+j} - \alpha_{n+j}) - \varepsilon_{n+j}\}\{Z_{n+j}(\bar{a}_{n+j} - \alpha_{n+j}) - \varepsilon_{n+j}\}'|Y_n]\\ &= Z_{n+j}\bar{P}_{n+j}Z_{n+j}' + H_{n+j}.\end{aligned}$$

We now derive recursions for calculating $\bar{a}_{n+j}$ and $\bar{P}_{n+j}$. We have $\alpha_{n+j+1} = T_{n+j}\alpha_{n+j} + R_{n+j}\eta_{n+j}$ so

$$\begin{aligned}\bar{a}_{n+j+1} &= T_{n+j}\,\mathrm{E}(\alpha_{n+j}|Y_n)\\ &= T_{n+j}\bar{a}_{n+j},\end{aligned}$$

for $j = 1, \dots, J-1$ and with $\bar{a}_{n+1} = a_{n+1}$. Also,

$$\begin{aligned}\bar{P}_{n+j+1} &= \mathrm{E}[(\bar{a}_{n+j+1} - \alpha_{n+j+1})(\bar{a}_{n+j+1} - \alpha_{n+j+1})'|Y_n]\\ &= T_{n+j}\,\mathrm{E}[(\bar{a}_{n+j} - \alpha_{n+j})(\bar{a}_{n+j} - \alpha_{n+j})'|Y_n]T'_{n+j}\\ &\quad + R_{n+j}\,\mathrm{E}[\eta_{n+j}\eta'_{n+j}]R'_{n+j}\\ &= T_{n+j}\bar{P}_{n+j}T'_{n+j} + R_{n+j}Q_{n+j}R'_{n+j},\end{aligned}$$

for $j = 1, \dots, J-1$.

We observe that the recursions for $\bar{a}_{n+j}$ and $\bar{P}_{n+j}$ are the same as the recursions for a_{n+j} and P_{n+j} of the Kalman filter (4.24) provided that we take $Z_{n+j} = 0$ for $j = 1, \dots, J-1$. But this is precisely the condition that in Section 4.10 enabled us to deal with missing observations by routine application of the Kalman filter. We have therefore demonstrated that forecasts of $y_{n+1}, \dots, y_{n+J}$ together with their forecast error variance matrices can be obtained merely by treating y_t for $t > n$ as missing observations and using the results of Section 4.10. In a sense this conclusion could be regarded as intuitively obvious; however, we thought it worthwhile demonstrating it algebraically. To sum up, forecasts and their associated error variance matrices can be obtained routinely in state space time series analysis based on linear Gaussian models by continuing the Kalman filter beyond $t = n$ with $Z_t = 0$ for $t > n$. Of course, for the computation of $\bar{y}_{n+j}$ and $\bar{F}_{n+j}$ we take Z_{n+j} as their actual values for $j = 1, \dots, J$. Similar results hold for forecasting values of the state vector α_t and hence for forecasting linear functions of elements of α_t. The results remain valid for MVLUE forecasting in the non-normal case and for Bayesian analysis using Lemmas 2 to 4. These results for forecasting are a particularly elegant feature of state space methods for time series analysis.

4.12 Dimensionality of observational vector

Throughout this chapter we have assumed, both for convenience of exposition and also because this is by far the most common case in practice, that the dimensionality of the observation vector y_t is a fixed value p. It is easy to verify, however, that none of the basic formulae that we have derived depend on this assumption. For example, the filtering recursion (4.24) and the disturbance smoother (4.69) both remain valid when the dimensionality of y_t is allowed to vary. This convenient generalisation arises because of the recursive nature of the formulae. In fact we made use of this property in the treatment of missing observational elements in Section 4.10. We do not, therefore, consider explicitly in this book the situation where the dimensionality of the state vector α_t varies with t, apart from the treatment of missing observations just referred to and the conversion of multivariate series to univariate series in Section 6.4.

4.13 Matrix formulations of basic results

In this section we provide matrix expressions for the state space model, filtering, smoothing and the associated unconditional and conditional densities. These expressions can give some additional insights into the filtering and smoothing results of this chapter. We further develop these expressions for reference purposes for the remaining chapters of this book, particularly in Part II.

4.13.1 State space model in matrix form

The linear Gaussian state space model (4.12) can itself be represented in a general matrix form. The observation equation can be formulated as

$$Y_n = Z\alpha + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, H), \tag{4.94}$$

with

$$Y_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad Z = \begin{bmatrix} Z_1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & Z_n & 0 \end{bmatrix}, \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \\ \alpha_{n+1} \end{pmatrix},$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \qquad H = \begin{bmatrix} H_1 & & 0 \\ & \ddots & \\ 0 & & H_n \end{bmatrix}. \tag{4.95}$$

The state equation takes the form

$$\alpha = T(\alpha_1^* + R\eta), \qquad \eta \sim \mathrm{N}(0, Q), \tag{4.96}$$

with

$$T = \begin{bmatrix} I & 0 & 0 & 0 & & 0 & 0 \\ T_1 & I & 0 & 0 & & 0 & 0 \\ T_2T_1 & T_2 & I & 0 & & 0 & 0 \\ T_3T_2T_1 & T_3T_2 & T_3 & I & & 0 & 0 \\ & & & & \ddots & & \vdots \\ T_{n-1}\cdots T_1 & T_{n-1}\cdots T_2 & T_{n-1}\cdots T_3 & T_{n-1}\cdots T_4 & & I & 0 \\ T_n\cdots T_1 & T_n\cdots T_2 & T_n\cdots T_3 & T_n\cdots T_4 & \cdots & T_n & I \end{bmatrix}, \tag{4.97}$$

$$\alpha_1^* = \begin{pmatrix} \alpha_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad R = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ R_1 & 0 & & 0 \\ 0 & R_2 & & 0 \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & R_n \end{bmatrix}, \tag{4.98}$$

$$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}, \qquad Q = \begin{bmatrix} Q_1 & & 0 \\ & \ddots & \\ 0 & & Q_n \end{bmatrix}. \tag{4.99}$$

This representation of the state space model is useful for getting a better understanding of some of the results in this chapter. For example, it follows that $\mathrm{E}(\alpha_1^*) = a_1^*$, $\mathrm{Var}(\alpha_1^*) = P_1^*$,

$$\begin{aligned} \mathrm{E}(\alpha) = Ta_1^*, \qquad \mathrm{Var}(\alpha) &= \mathrm{Var}\left\{T\left[(\alpha_1^* - a_1^*) + R\eta\right]\right\} \\ &= T(P_1^* + RQR')T' \\ &= TQ^*T', \end{aligned} \tag{4.100}$$

where

$$a_1^* = \begin{pmatrix} a_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad P_1^* = \begin{bmatrix} P_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & & 0 \end{bmatrix}, \qquad Q^* = P_1^* + RQR'.$$

Furthermore, we show that the observation vectors y_t are linear functions of the initial state vector α_1 and the disturbance vectors ε_t and η_t for $t = 1, \dots, n$ since it follows by substitution of (4.96) into (4.94) that

$$Y_n = ZT\alpha_1^* + ZTR\eta + \varepsilon. \tag{4.101}$$

It also follows that

$$\begin{aligned} \mathrm{E}(Y_n) = \mu = ZTa_1^*, \qquad \mathrm{Var}(Y_n) = \Omega &= ZT(P_1^* + RQR')T'Z' + H \\ &= ZTQ^*T'Z' + H. \end{aligned} \tag{4.102}$$

4.13.2 Matrix expression for densities

Given the model expressions in matrix form, we can also express the densities of $p(Y_n)$ and $p(\alpha, Y_n)$ in terms of vectors and matrices. For example, it follows from (4.102) that the log of the density function $p(Y_n)$ is given by

$$\log p(Y_n) = \text{constant} - \frac{1}{2}\log|\,\Omega\,| - \frac{1}{2}(Y_n - \mu)'\Omega^{-1}(Y_n - \mu). \tag{4.103}$$

An expression for the joint density of Y_n and α can be based on the decomposition $p(\alpha, Y_n) = p(Y_n|\alpha)p(\alpha)$. It follows from (4.100) that the logdensity of α is given by

$$\log p(\alpha) = \text{constant} - \frac{1}{2}\log|V^*| - \frac{1}{2}(\alpha - a^*)' V^{*\,-1}(\alpha - a^*), \tag{4.104}$$

where $a^* = \mathrm{E}(\alpha) = Ta_1^*$ and $V^* = \mathrm{Var}(\alpha) = TQ^*T'$. The observation equation (4.94) implies that the logdensity of the observation vector Y_n given the state α is given by

$$\log p(Y_n|\alpha) = \text{constant} - \frac{1}{2}\log|H| - \frac{1}{2}(Y_n - \theta)' H^{-1}(Y_n - \theta), \tag{4.105}$$

where $\theta = Z\alpha$ is referred to as the signal. It follows that

$$p(Y_n|\alpha) = p(Y_n|\theta).$$

The joint logdensity $\log p(\alpha, Y_n)$ is simply the sum of $\log p(\alpha)$ and $\log p(Y_n|\theta)$.

4.13.3 Filtering in matrix form: Cholesky decomposition

We now show that the vector of innovations can be represented as $v = CY_n - Ba_1^*$ where $v = (v_1', \ldots, v_n')'$ and where C and B are matrices of which C is lower block triangular. First we observe that

$$a_{t+1} = L_t a_t + K_t y_t,$$

which follows from (4.21) with $v_t = y_t - Z_t a_t$ and $L_t = T_t - K_t Z_t$. Then by substituting repeatedly we have

$$a_{t+1} = L_t L_{t-1} \cdots L_1 a_1 + \sum_{j=1}^{t-1} L_t L_{t-1} \cdots L_{j+1} K_j y_j + K_t y_t$$

and

$$\begin{aligned} v_1 &= y_1 - Z_1 a_1, \\ v_2 &= -Z_2 L_1 a_1 + y_2 - Z_2 K_1 y_1, \\ v_3 &= -Z_3 L_2 L_1 a_1 + y_3 - Z_3 K_2 y_2 - Z_3 L_2 K_1 y_1, \end{aligned}$$

and so on. Generally,

$$\begin{aligned} v_t = &-Z_t L_{t-1} L_{t-2} \cdots L_1 a_1 + y_t - Z_t K_{t-1} y_{t-1} \\ &- Z_t \sum_{j=1}^{t-2} L_{t-1} \cdots L_{j+1} K_j y_j. \end{aligned}$$

Note that the matrices K_t and L_t depend on P_1, Z, T, R, H and Q but do not depend on the initial mean vector a_1 or the observations $y_1, \dots, y_n$, for $t = 1, \dots, n$. The innovations can thus be represented as

$$\begin{aligned} v &= (I - ZLK)Y_n - ZLa_1^* \\ &= CY_n - Ba_1^*, \end{aligned} \tag{4.106}$$

where $C = I - ZLK$, $B = ZL$,

$$L = \begin{bmatrix} I & 0 & 0 & 0 & & 0 & 0 \\ L_1 & I & 0 & 0 & & 0 & 0 \\ L_2L_1 & L_2 & I & 0 & & 0 & 0 \\ L_3L_2L_1 & L_3L_2 & L_3 & I & & 0 & 0 \\ & & & & \ddots & & \vdots \\ L_{n-1}\cdots L_1 & L_{n-1}\cdots L_2 & L_{n-1}\cdots L_3 & L_{n-1}\cdots L_4 & & I & 0 \\ L_n\cdots L_1 & L_n\cdots L_2 & L_n\cdots L_3 & L_n\cdots L_4 & \cdots & L_n & I \end{bmatrix},$$

$$K = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ K_1 & 0 & \cdots & 0 \\ 0 & K_2 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & K_n \end{bmatrix},$$

and matrix Z is defined in (4.95). It can be easily verified that matrix C is lower block triangular with identity matrices on the leading diagonal blocks. Since $v = CY_n - ZLa_1^*$, $\mathrm{Var}(Y_n) = \Omega$ and a_1^* is constant, it follows that $\mathrm{Var}(v) = C\Omega C'$. However, we know from Subsection 4.3.5 that the innovations are independent of each other so that $\mathrm{Var}(v)$ is the block diagonal matrix

$$F = \begin{bmatrix} F_1 & 0 & \cdots & 0 \\ 0 & F_2 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & F_n \end{bmatrix}.$$

This shows that in effect the Kalman filter is essentially equivalent to a block version of a Cholesky decomposition applied to the observational variance matrix implied by the state space model (4.12). A general discussion of the Cholesky decomposition is provided by Golub and Van Loan (1996, §4.2).

Given the special structure of $C = I - ZLK$ we can reproduce some interesting results in matrix form. We first notice from (4.106) that $\mathrm{E}(v) = C\mathrm{E}(Y_n) - ZLa_1^* = 0$. Since $\mathrm{E}(Y_n) = \mu = ZTa_1^*$ from (4.102), we obtain the identity $CZT = ZL$. It follows that

$$v = C(Y_n - \mu), \qquad F = \mathrm{Var}(v) = C\Omega C'. \tag{4.107}$$

Further we notice that C is nonsingular and

$$\Omega^{-1} = C' F^{-1} C. \tag{4.108}$$

It can be verified that matrix $C = I - ZLK$ is a lower triangular block matrix with its leading diagonal blocks equal to identity matrices. It follows that $|\,C\,| = 1$ and from (4.108) that $|\,\Omega\,|^{-1} = |\,\Omega^{-1}\,| = |\,C'F^{-1}C\,| = |\,C\,|\cdot|\,F\,|^{-1}\cdot|\,C\,| = |\,F\,|^{-1}$. This result is particularly useful for the evaluation of the log of the density $p(Y_n)$ in (4.103). By applying the Cholesky decomposition to (4.103), we obtain

$$\log p(Y_n) = \text{constant} - \frac{1}{2}\log|\,F\,| - \frac{1}{2}v'F^{-1}v, \tag{4.109}$$

which follows directly from (4.107) and (4.108). The Kalman filter computes v and F in a computationally efficient way and is therefore instrumental for the evaluation of (4.109).

4.13.4 Smoothing in matrix form

Let $\hat\varepsilon = (\hat\varepsilon_1', \dots, \hat\varepsilon_n')'$ where $\hat\varepsilon_t = \mathrm{E}(\varepsilon_t|Y_n)$, for $t = 1, \dots, n$, is evaluated as described in Subsection 4.5.3. The smoothed observation disturbance vector $\hat\varepsilon$ can be obtained directly via the application of Lemma 1, that is

$$\hat\varepsilon = \mathrm{E}(\varepsilon|Y_n) = \mathrm{Cov}(\varepsilon, Y_n)\Omega^{-1}(Y_n - \mu).$$

Since $\mathrm{Cov}(\varepsilon, Y_n) = H$, it follows by the substitution of (4.107) and (4.108) that

$$\begin{aligned}\hat\varepsilon &= H\Omega^{-1}(Y_n - \mu)\\ &= Hu,\end{aligned}$$

where

$$\begin{aligned}u &= \Omega^{-1}(Y_n - \mu)\\ &= C'F^{-1}v\\ &= (I - K'L'Z')F^{-1}v\\ &= F^{-1}v - K'r,\end{aligned}$$

with $r = L'Z'F^{-1}v$. It is easily verified that the definitions of u and r are consistent with the definitions of their elements in (4.59) and (4.38), respectively.

Let $\hat\eta = (\hat\eta_1', \dots, \hat\eta_n')'$ where $\hat\eta_t = \mathrm{E}(\eta_t|Y_n)$, for $t = 1, \dots, n$, is evaluated as described in Subsection 4.5.3. We obtain the stack of smoothed state disturbance vector $\hat\eta$ directly via

$$\begin{aligned}\hat\eta &= \mathrm{Cov}(\eta, Y_n)\Omega^{-1}(Y_n - \mu)\\ &= QR'T'Z'u\\ &= QR'r,\end{aligned}$$

where $r = T'Z'u = T'Z'\Omega^{-1}(Y_n - \mu) = L'Z'F^{-1}v$ since $CZT = ZL$. This result is consistent with the definitions of the elements of $\hat{\eta}$, that is $\hat{\eta}_t = Q_t R_t' r_t$ where r_t is evaluated by $r_{t-1} = Z_t' u_t + T_t' r_t$; see Subsection 4.5.3.

Finally, we obtain the smoothed estimator of α via

$$\begin{aligned}\hat{\alpha} &= \mathrm{E}(\alpha) + \mathrm{Cov}(\alpha, Y_n)\Omega^{-1}(Y_n - \mu) \\ &= \mathrm{E}(\alpha) + \mathrm{Cov}(\alpha, Y_n)u \\ &= Ta_1^* + TQ^*T'Z'u \\ &= Ta_1^* + TQ^*r, \end{aligned} \tag{4.110}$$

since $Y_n = Z\alpha + \varepsilon$ and $\mathrm{Cov}(\alpha, Y_n) = \mathrm{Var}(\alpha)Z' = TQ^*T'Z'$. This is consistent with the way $\hat{\alpha}_t$ is evaluated using fast state smoothing as described in Subsection 4.6.2.

4.13.5 Matrix expressions for signal

Given equation (4.102) and the definition of the signal $\theta = Z\alpha$, we further define

$$\mu = \mathrm{E}(\theta) = \mathrm{E}(Z\alpha) = Za^* = ZTa_1^*, \qquad \Psi = \mathrm{Var}(\theta) = ZV^*Z' = ZTQ^*T'Z'.$$

The logdensity of the signal is therefore given by

$$\log p(\theta) = \text{constant} - \frac{1}{2}\log|\Psi| - \frac{1}{2}(\theta - \mu)'\Psi^{-1}(\theta - \mu). \tag{4.111}$$

Also from equation (4.102) we have

$$\mathrm{E}(Y_n) = \mu, \qquad \mathrm{Var}(Y_n) = \Omega = \Psi + H.$$

Since $\mathrm{Cov}(\theta, Y_n) = \mathrm{Var}(\theta) = \Psi$, it follows from Lemma 1 that the conditional (smoothed) mean and variance of the signal is given by

$$\hat{\theta} = \mathrm{E}(\theta|Y_n) = \mu + \Psi\Omega^{-1}(Y_n - \mu), \qquad \mathrm{Var}(\theta|Y_n) = \Psi - \Psi\Omega^{-1}\Psi. \tag{4.112}$$

In the case of the smoothed mean $\hat{\theta}$, after some matrix manupulation, we obtain

$$\hat{\theta} = (\Psi^{-1} + H^{-1})^{-1}(\Psi^{-1}\mu + H^{-1}Y_n).$$

It should be kept in mind that this expression is computed by the Kalman filter and smoothing recursions. In particular, the application of Kalman filter and disturbance smoother is sufficient since $\hat{\theta} = Y_n - \hat{\varepsilon}$. In the linear Gaussian model (4.12) all random variables are Gaussian and all relations between the variables are linear. Therefore, the mean and the mode are equal. It follows that $\hat{\theta}$ is also the mode of the smoothed logdensity $p(\theta|Y_n)$.

The expression for the smoothed signal $\hat{\theta} = \mathrm{E}(\theta|Y_n)$ leads to an interesting result for the computation of u. It follows from (4.112) that

$$u = \Omega^{-1}(Y_n - \mu) = \Psi^{-1}(\hat{\theta} - \mu). \tag{4.113}$$

Since $\Omega = \Psi + H$, the expression (4.113) implies that u can also be computed by applying the Kalman filter and smoother to a linear Gaussian state space model for $\hat{\theta}$ without observation noise such that $H = 0$ and $\Omega = \Psi$, that is

$$\hat{\theta}_t = Z_t \alpha_t, \qquad \alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \tag{4.114}$$

for $t = 1, \ldots, n$. For example, result (4.113) implies that $\hat{\alpha} = \mathrm{E}(\alpha|Y_n)$ can be computed from applying the Kalman filter and smoother to model (4.114) since

$$\hat{\alpha} = Ta_1^* + TQ^*T'Z'u = Ta_1^* + TQ^*T'Z'\Psi^{-1}(\hat{\theta} - \mu).$$

For specific applications these results can be exploited to improve computational efficiency.

4.13.6 Simulation smoothing

Given the results discussed in Section 4.9 together with the matrix expressions and results in this section, we can develop convenient expressions for simulation smoothing for signal and state vectors. In the case of simulation smoothing for the signal, it follows from the discussion in Subsection 4.9.2 that we can express the draw from $p(\theta|Y_n)$ by

$$\tilde{\theta} = \theta^+ - \hat{\theta}^+ + \hat{\theta},$$

where

$$\theta^+ \sim p(\theta), \qquad \hat{\theta}^+ = \mathrm{E}(\theta|y^+), \qquad \hat{\theta} = \mathrm{E}(\theta|Y_n),$$

with

$$y^+ = \theta^+ + \varepsilon^+, \qquad \theta^+ = Z\alpha^+ = ZT(\alpha_1^{*\,+} + R\eta^+),$$

and

$$\varepsilon^+ \sim \mathrm{N}(0, H), \qquad \alpha_1^+ \sim \mathrm{N}(a, P), \qquad \eta^+ \sim \mathrm{N}(0, Q).$$

We note that $\alpha_1^{*\,+\,\prime} = (\alpha_1^{+\,\prime}, 0, \ldots, 0)'$. Simulation smooting reduces the application of a single Kalman filter and smoother since

$$\tilde{\theta} - \theta^+ = \hat{\theta} - \hat{\theta}^+ = [\mu + \Psi\Omega^{-1}(Y_n - \mu)] - [\mu + \Psi\Omega^{-1}(y^+ - \mu)] = \Psi\Omega^{-1}(Y_n - y^+),$$

using (4.112). It follows that

$$\tilde{\theta} = \theta^+ + \Psi\Omega^{-1}(Y_n - y^+).$$

Once α^+, $\theta^+ = Z\alpha^+$ and $y^+ = \theta^+ + \varepsilon^+$ are computed using the relations in the linear Gaussian state space model (4.12), the sample $\tilde{\theta} \sim p(\theta|Y_n)$ is obtained from the Kalman filter and smoother applied to model (4.12) with $a_1 = 0$ and for the 'observations' $y_t - y_t^+$, for $t = 1, \ldots, n$. Similar arguments apply to the computation of $\tilde{\alpha}$, $\tilde{\varepsilon}$ and $\tilde{\eta}$.

4.14 Exercises

4.14.1

Taking the notation of Section 4.2 and

$$\Sigma_* = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma'_{xy} & \Sigma_{yy} \end{bmatrix},$$

verify that

$$\Sigma_* = \begin{bmatrix} I & \Sigma_{xy}\Sigma_{yy}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma'_{xy} & 0 \\ 0 & \Sigma_{yy} \end{bmatrix} \begin{bmatrix} I & 0 \\ \Sigma_{yy}^{-1}\Sigma'_{xy} & I \end{bmatrix},$$

and hence that

$$\Sigma_*^{-1} = \begin{bmatrix} I & 0 \\ -\Sigma_{yy}^{-1}\Sigma'_{xy} & I \end{bmatrix} \begin{bmatrix} \left(\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma'_{xy}\right)^{-1} & 0 \\ 0 & \Sigma_{yy}^{-1} \end{bmatrix} \begin{bmatrix} I & -\Sigma_{xy}\Sigma_{yy}^{-1} \\ 0 & I \end{bmatrix}.$$

Taking

$$p(x, y) = \text{constant} \times \exp\left[-\frac{1}{2}\begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix}\right],$$

obtain $p(x|y)$ and hence prove Lemma 1. This exercise contains the essentials of the proofs of our Lemma 1 in Anderson and Moore (1979, Example 3.2) and Harvey (1989, Appendix to Chapter 3).

4.14.2

Under the conditions of Lemma 1 and using the expression for Σ^{-1} in Exercise 4.14.1, show that $\widehat{x} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$ is the maximum likelihood estimator of x for given y with asymptotic variance $\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma'_{xy}$.

4.14.3

Suppose that the fixed vector λ is regarded as an estimate of a random vector x whose mean vector is μ. Show that the minimum mean square error matrix $\mathrm{E}\left[(\lambda - x)(\lambda - x)'\right]$ is obtained when $\lambda = \mu$.

4.14.4

In the joint distribution, not necessarily normal, of random vectors x and y, suppose that $\bar{x} = \beta + \gamma y$ is an estimate of x given y with mean square error matrix $\mathrm{MSE}(\bar{x}) = \mathrm{E}\left[(\bar{x} - x)(\bar{x} - x)'\right]$. Using Exercise 4.14.3 and details of the proof of Lemma 2, show that the minimum mean square error matrix is obtained when $\bar{x} = \widehat{x}$ where $\widehat{x}$ is defined by (4.6) with $\mathrm{MSE}(\widehat{x})$ given by (4.7).

4.14.5

Adopt the notation of Lemma 3 and assume that $p(y|x) = \mathrm{N}(Zx, H)$, where Z and H are constant matrices of appropriate dimensions.

(a) Show that

$$p(x|y) = \exp(-\frac{1}{2}Q)$$

where

$$\begin{aligned} Q &= x'(Z'H^{-1}Z + \Sigma_{yy}^{-1})x - 2(y'H^{-1}Z + \mu_y'\Sigma_{yy}^{-1})x + \text{constant} \\ &= (x-m)'C^{-1}(x-m) + \text{constant}, \end{aligned}$$

with

$$C^{-1} = \Sigma_{yy}^{-1} + Z'H^{-1}Z, \qquad m = C(Z'H^{-1}y + \Sigma_{yy}^{-1}\mu_y).$$

(b) Using the matrix identity

$$(\Sigma_{yy}^{-1} + Z'H^{-1}Z)^{-1} = \Sigma_{yy} - \Sigma_{yy}Z'(Z\Sigma_{yy}Z' + H)^{-1}Z\Sigma_{yy},$$

show that the result in (a) proves Lemma 3 for this form of $p(x|y)$.

This exercise contains the essentials of the proof in West and Harrison (1997, §17.3.3) of our Lemma 1.

4.14.6

Derive the Kalman filter equations of Subsection 4.3.1 in case $\mathrm{E}(\varepsilon_t\eta_t') = R_t^*$ and $\mathrm{E}(\varepsilon_t\eta_s') = 0$ where R_t^* is a fixed and known $p \times r$ matrix for $t, s = 1, \ldots, n$ and $t \neq s$.

4.14.7

Given the state space model (4.12) and the results in Sections 4.4 and 4.5, derive recursive expressions for

$$\mathrm{Cov}(\varepsilon_t, \alpha_t|Y_n), \qquad \mathrm{Cov}(\eta_t, \alpha_t|Y_n),$$

for $t = 1, \ldots, n$.

4.14.8

How would you modify the state space model to carry out fixed-lag smoothing when you want to rely only on the smoothing recursions (4.44) in Subsection 4.4.4? See also Exercise 4.14.7.

5 Initialisation of filter and smoother

5.1 Introduction

In the previous chapter we have considered the operations of filtering and smoothing for the linear Gaussian state space model

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{N}(0, H_t), \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t, & \eta_t &\sim \mathrm{N}(0, Q_t), \end{aligned} \tag{5.1}$$

under the assumption that $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known. In most practical applications, however, at least some of the elements of a_1 and P_1 are unknown. We now develop methods of starting up the series when this is the situation; the process is called *initialisation*. We shall consider the general case where some elements of α_1 have a known joint distribution while about other elements we are completely ignorant. We treat the case in detail where the observations are normally distributed from a classical point of view. The results can be extended to minimum variance unbiased linear estimates and to Bayesian analysis by Lemmas 2, 3 and 4.

A general model for the initial state vector α_1 is

$$\alpha_1 = a + A\delta + R_0\eta_0, \qquad \eta_0 \sim \mathrm{N}(0, Q_0), \tag{5.2}$$

where the $m \times 1$ vector a is known, δ is a $q \times 1$ vector of unknown quantities, the $m \times q$ matrix A and the $m \times (m - q)$ matrix R_0 are selection matrices, that is, they consist of columns of the identity matrix I_m; they are defined so that when taken together, their columns constitute a set of g columns of I_m with $g \leq m$ and $A'R_0 = 0$. The matrix Q_0 is assumed to be positive definite and known. In most cases vector a will be treated as a zero vector unless some elements of the initial state vector are known constants. When all elements of the state vector α_t are stationary, the initial means, variances and covariances of these initial state elements can be derived from the model parameters. For example, in the case of a stationary ARMA model it is straightforward to obtain the unconditional variance matrix Q_0 as we will show in Subsection 5.6.2. The Kalman filter (4.24) can then be applied routinely with $a_1 = 0$ and $P_1 = Q_0$.

To illustrate the structure and notation of (5.2) for readers unfamiliar with the subject, we present a simple example in which

$$y_t = \mu_t + \rho_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}\big(0, \sigma^2_\varepsilon\big),$$

where

$$\begin{aligned} \mu_{t+1} &= \mu_t + \nu_t + \xi_t, \qquad & \xi_t &\sim \mathrm{N}\big(0, \sigma_\xi^2\big), \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim \mathrm{N}\big(0, \sigma_\zeta^2\big), \\ \rho_{t+1} &= \phi\rho_t + \tau_t, & \tau_t &\sim \mathrm{N}\big(0, \sigma_\tau^2\big), \end{aligned}$$

in which $|\phi| < 1$ and the disturbances are all mutually and serially uncorrelated. Thus μ_t is a local linear trend as in (3.2), which is nonstationary, while ρ_t is an unobserved stationary first order AR(1) series with zero mean. In state space form this is

$$y_t = (1 \quad 0 \quad 1)\begin{pmatrix} \mu_t \\ \nu_t \\ \rho_t \end{pmatrix} + \varepsilon_t,$$

$$\begin{pmatrix} \mu_{t+1} \\ \nu_{t+1} \\ \rho_{t+1} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \phi \end{bmatrix}\begin{pmatrix} \mu_t \\ \nu_t \\ \rho_t \end{pmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{pmatrix} \xi_t \\ \zeta_t \\ \tau_t \end{pmatrix}.$$

Thus we have

$$a = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad R_0 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

with $\eta_0 = \rho_1$ and where $Q_0 = \sigma_\tau^2/(1 - \phi^2)$ is the variance of the stationary series ρ_t.

Although we treat the parameter ϕ as known for the purpose of this section, in practice it is unknown, which in a classical analysis is replaced by its maximum likelihood estimate. We see that the object of the representation (5.2) is to separate out α_1 into a constant part a, a nonstationary part $A\delta$ and a stationary part $R_0\eta_0$. In a Bayesian analysis, α_1 can be treated as having a known or noninformative prior density.

The vector δ can be treated as a fixed vector of unknown parameters or as a vector of random normal variables with infinite variances. For the case where δ is fixed and unknown, we may estimate it by maximum likelihood; a technique for doing this was developed by Rosenberg (1973) and we will discuss this in Section 5.7. For the case where δ is random we assume that

$$\delta \sim \mathrm{N}(0, \kappa I_q), \tag{5.3}$$

where we let $\kappa \to \infty$. We begin by considering the Kalman filter with initial conditions $a_1 = \mathrm{E}(\alpha_1) = a$ and $P_1 = \mathrm{Var}(\alpha_1)$ where

$$P_1 = \kappa P_\infty + P_*, \tag{5.4}$$

and we let $\kappa \to \infty$ at a suitable point later. Here $P_\infty = AA'$ and $P_* = R_0 Q_0 R_0'$; since A consists of columns of I_m it follows that P_∞ is an $m \times m$ diagonal matrix with q diagonal elements equal to one and the other elements equal to zero. Also, without loss of generality, when a diagonal element of P_∞ is nonzero we take the corresponding element of a to be zero. A vector δ with distribution $\mathrm{N}(0, \kappa I_q)$ as $\kappa \to \infty$ is said to be *diffuse*. Initialisation of the Kalman filter when some elements of α_1 are diffuse is called *diffuse initialisation* of the filter. We now consider the modifications required to the Kalman filter in the diffuse initialisation case.

A simple approximate technique is to replace κ in (5.4) by an arbitrary large number and then use the standard Kalman filter (4.13). This approach was employed by Harvey and Phillips (1979). While the device can be useful for approximate exploratory work, it is not recommended for general use since it can lead to large rounding errors. We therefore develop an exact treatment.

The technique we shall use is to expand matrix products as power series in κ^{-1}, taking only the first two or three terms as required, and then let $\kappa \to \infty$ to obtain the dominant term. The underlying idea was introduced by Ansley and Kohn (1985) in a somewhat inaccessible way. Koopman (1997) presented a more transparent treatment of diffuse filtering and smoothing based on the same idea. Further developments were given by Koopman and Durbin (2003) who obtained the results that form the basis of Section 5.2 on filtering and Section 5.3 on state smoothing. This approach gives recursions different from those obtained from the augmentation technique of de Jong (1991) which is based on ideas of Rosenberg (1973); see Section 5.7. Illustrations of these initialisation methods are given in Section 5.6 and Subsection 5.7.4.

A direct approach to the initialisation problem for the general multivariate linear Gaussian state space model turns out to be somewhat complicated as can be seen from the treatment of Koopman (1997). The reason for this is that for multivariate series the inverse matrix F_t^{-1} does not have a simple general expansion in powers of κ^{-1} for the first few terms of the series, due to the fact that in very specific situations the part of F_t associated with P_∞ can be singular with varying rank. Rank deficiencies may occur when observations are missing near the beginning of the series, for example. For univariate series, however, the treatment is much simpler since F_t is a scalar so the part associated with P_∞ can only be either zero or positive, both of which are easily dealt with. In complicated cases, it turns out to be simpler in the multivariate case to adopt the filtering and smoothing approach of Section 6.4 in which the multivariate series is converted to a univariate series by introducing the elements of the observational vector y_t one at a time, rather than dealing with the series directly as a multivariate series. We therefore begin by assuming that the part of F_t associated with P_∞ is nonsingular or zero for any t. In this way we can treat most multivariate series, for which this assumption holds directly, and at the same time obtain general results for all univariate time series. We shall use these results in Section 6.4 for the univariate treatment of multivariate series.

5.2 The exact initial Kalman filter

In this section we use the notation $O(\kappa^{-j})$ to denote a function $f(\kappa)$ of κ such that the limit of $\kappa^j f(\kappa)$ as $\kappa \to \infty$ is finite for $j = 1, 2$.

5.2.1 The basic recursions

Analogously to the decomposition of the initial matrix P_1 in (5.4) we show that the mean square error matrix P_t has the decomposition

$$P_t = \kappa P_{\infty,t} + P_{*,t} + O(\kappa^{-1}), \qquad t = 2, \ldots, n, \tag{5.5}$$

where $P_{\infty,t}$ and $P_{*,t}$ do not depend on κ. It will be shown that $P_{\infty,t} = 0$ for $t > d$ where d is a positive integer which in normal circumstances is small relative to n. The consequence is that the usual Kalman filter (4.24) applies without change for $t = d + 1, \ldots, n$ with $P_t = P_{*,t}$. Note that when all initial state elements have a known joint distribution or are fixed and known, matrix $P_\infty = 0$ and therefore $d = 0$.

The decomposition (5.5) leads to the similar decompositions

$$F_t = \kappa F_{\infty,t} + F_{*,t} + O(\kappa^{-1}), \qquad M_t = \kappa M_{\infty,t} + M_{*,t} + O(\kappa^{-1}), \tag{5.6}$$

and, since $F_t = Z_t P_t Z_t' + H_t$ and $M_t = P_t Z_t'$, we have

$$\begin{aligned} F_{\infty,t} &= Z_t P_{\infty,t} Z_t', \qquad & F_{*,t} &= Z_t P_{*,t} Z_t' + H_t, \\ M_{\infty,t} &= P_{\infty,t} Z_t', \qquad & M_{*,t} &= P_{*,t} Z_t', \end{aligned} \tag{5.7}$$

for $t = 1, \ldots, d$. The Kalman filter that we shall derive as $\kappa \to \infty$ we shall call the *exact initial Kalman filter*. We use the word *exact* here to distinguish the resulting filter from the approximate filter obtained by choosing an arbitrary large value for κ and applying the standard Kalman filter (4.24). In deriving it, it is important to note from (5.7) that a zero matrix $M_{\infty,t}$ (whether $P_{\infty,t}$ is a zero matrix or not) implies that $F_{\infty,t} = 0$. As in the development of the Kalman filter in Subsection 4.3.2 we assume that F_t is nonsingular. The derivation of the exact initial Kalman filter is based on the expansion for $F_t^{-1} = [\kappa F_{\infty,t} + F_{*,t} + O(\kappa^{-1})]^{-1}$ as a power series in κ^{-1}, that is

$$F_t^{-1} = F_t^{(0)} + \kappa^{-1} F_t^{(1)} + \kappa^{-2} F_t^{(2)} + O(\kappa^{-3}), \tag{5.8}$$

for large κ. Since $I_p = F_t F_t^{-1}$ we have

$$\begin{aligned} I_p &= (\kappa F_{\infty,t} + F_{*,t} + \kappa^{-1} F_{a,t} + \kappa^{-2} F_{b,t} + \cdots) \\ &\quad \times \left(F_t^{(0)} + \kappa^{-1} F_t^{(1)} + \kappa^{-2} F_t^{(2)} + \cdots\right). \end{aligned}$$

On equating coefficients of κ^j for $j = 0, -1, -2, \ldots$ we obtain

$$\begin{aligned} F_{\infty,t}F_t^{(0)} &= 0, \\ F_{*,t}F_t^{(0)} + F_{\infty,t}F_t^{(1)} &= I_p, \\ F_{a,t}F_t^{(0)} + F_{*,t}F_t^{(1)} + F_{\infty,t}F_t^{(2)} &= 0, \quad \text{etc.} \end{aligned} \tag{5.9}$$

We need to solve equations (5.9) for $F_t^{(0)}$, $F_t^{(1)}$ and $F_t^{(2)}$; further terms are not required. We shall consider only the cases where $F_{\infty,t}$ is nonsingular or $F_{\infty,t} = 0$. This limitation of the treatment is justified for three reasons. First, it gives a complete solution for the important special case of univariate series, since if y_t is univariate $F_{\infty,t}$ must obviously be positive or zero. Second, if y_t is multivariate the restriction is satisfied in most practical cases. Third, for those rare cases where y_t is multivariate but the restriction is not satisfied, the series can be dealt with as a univariate series by the technique described in Section 6.4. By limiting the treatment at this point to these two cases, the derivations are essentially no more difficult than those required for treating the univariate case. However, solutions for the general case where no restrictions are placed on $F_{\infty,t}$ are algebraically complicated; see Koopman (1997). We mention that although $F_{\infty,t}$ nonsingular is the most common case, situations can arise in practice where $F_{\infty,t} = 0$ even when $P_{\infty,t} \neq 0$ if $M_{\infty,t} = P_{\infty,t}Z_t' = 0$.

Taking first the case where $F_{\infty,t}$ is nonsingular we have from (5.9),

$$F_t^{(0)} = 0, \qquad F_t^{(1)} = F_{\infty,t}^{-1}, \qquad F_t^{(2)} = -F_{\infty,t}^{-1}F_{*,t}F_{\infty,t}^{-1}. \tag{5.10}$$

The matrices $K_t = T_tM_tF_t^{-1}$ and $L_t = T_t - K_tZ_t$ depend on the inverse matrix F_t^{-1} so they also can be expressed as power series in κ^{-1}. We have

$$K_t = T_t[\kappa M_{\infty,t} + M_{*,t} + O(\kappa^{-1})]\big(\kappa^{-1}F_t^{(1)} + \kappa^{-2}F_t^{(2)} + \cdots\big),$$

so

$$K_t = K_t^{(0)} + \kappa^{-1}K_t^{(1)} + O(\kappa^{-2}), \qquad L_t = L_t^{(0)} + \kappa^{-1}L_t^{(1)} + O(\kappa^{-2}), \tag{5.11}$$

where

$$\begin{aligned} K_t^{(0)} &= T_tM_{\infty,t}F_t^{(1)}, & L_t^{(0)} &= T_t - K_t^{(0)}Z_t, \\ K_t^{(1)} &= T_tM_{*,t}F_t^{(1)} + T_tM_{\infty,t}F_t^{(2)}, & L_t^{(1)} &= -K_t^{(1)}Z_t. \end{aligned} \tag{5.12}$$

By following the recursion (4.21) for a_{t+1} starting with $t = 1$ we find that a_t has the form

$$a_t = a_t^{(0)} + \kappa^{-1}a_t^{(1)} + O(\kappa^{-2}),$$

where $a_1^{(0)} = a$ and $a_1^{(1)} = 0$. As a consequence v_t has the form

$$v_t = v_t^{(0)} + \kappa^{-1}v_t^{(1)} + O(\kappa^{-2}),$$

where $v_t^{(0)} = y_t - Z_t a_t^{(0)}$ and $v_t^{(1)} = -Z_t a_t^{(1)}$. The updating equation (4.21) for a_{t+1} can now be expressed as

$$\begin{aligned} a_{t+1} &= T_t a_t + K_t v_t \\ &= T_t\big[a_t^{(0)} + \kappa^{-1} a_t^{(1)} + O(\kappa^{-2})\big] \\ &\quad + \big[K_t^{(0)} + \kappa^{-1} K_t^{(1)} + O(\kappa^{-2})\big]\big[v_t^{(0)} + \kappa^{-1} v_t^{(1)} + O(\kappa^{-2})\big], \end{aligned}$$

which becomes as $\kappa \to \infty$,

$$a_{t+1}^{(0)} = T_t a_t^{(0)} + K_t^{(0)} v_t^{(0)}, \qquad t = 1, \dots, n. \tag{5.13}$$

The updating equation (4.23) for P_{t+1} is

$$\begin{aligned} P_{t+1} &= T_t P_t L_t' + R_t Q_t R_t' \\ &= T_t[\kappa P_{\infty,t} + P_{*,t} + O(\kappa^{-1})]\big[L_t^{(0)} + \kappa^{-1} L_t^{(1)} + O(\kappa^{-2})\big]' + R_t Q_t R_t'. \end{aligned}$$

Consequently, on letting $\kappa \to \infty$, the updates for $P_{\infty,t+1}$ and $P_{*,t+1}$ are given by

$$\begin{aligned} P_{\infty,t+1} &= T_t P_{\infty,t} L_t^{(0)\prime}, \\ P_{*,t+1} &= T_t P_{\infty,t} L_t^{(1)\prime} + T_t P_{*,t} L_t^{(0)\prime} + R_t Q_t R_t', \end{aligned} \tag{5.14}$$

for $t = 1, \dots, n$. The matrix P_{t+1} also depends on terms in κ^{-1}, κ^{-2}, etc. but these terms will not be multiplied by κ or higher powers of κ within the Kalman filter recursions. Thus the updating equations for P_{t+1} do not need to take account of these terms. Recursions (5.13) and (5.14) constitute the exact Kalman filter.

In the case where $F_{\infty,t} = 0$, we have

$$F_t = F_{*,t} + O(\kappa^{-1}), \qquad M_t = M_{*,t} + O(\kappa^{-1}),$$

and the inverse matrix F_t^{-1} is given by

$$F_t^{-1} = F_{*,t}^{-1} + O(\kappa^{-1}).$$

Therefore,

$$\begin{aligned} K_t &= T_t[M_{*,t} + O(\kappa^{-1})]\big[F_{*,t}^{-1} + O(\kappa^{-1})\big] \\ &= T_t M_{*,t} F_{*,t}^{-1} + O(\kappa^{-1}). \end{aligned}$$

The updating equation for $a_{t+1}^{(0)}$ (5.13) has

$$K_t^{(0)} = T_t M_{*,t} F_{*,t}^{-1}, \tag{5.15}$$

and the updating equation for P_{t+1} becomes

$$\begin{aligned}P_{t+1} &= T_t P_t L_t' + R_t Q_t R_t' \\ &= T_t[\kappa P_{\infty,t} + P_{*,t} + O(\kappa^{-1})]\big[L_t^{(0)} + \kappa^{-1}L_t^{(1)} + O(\kappa^{-2})\big]' + R_t Q_t R_t',\end{aligned}$$

where $L_t^{(0)} = T_t - K_t^{(0)} Z_t$ and $L_t^{(1)} = -K_t^{(1)} Z_t$. The updates for $P_{\infty,t+1}$ and $P_{*,t}$ can be simplified considerably since $M_{\infty,t} = P_{\infty,t} Z_t' = 0$ when $F_{\infty,t} = 0$. By letting $\kappa \to \infty$ we have

$$\begin{aligned}P_{\infty,t+1} &= T_t P_{\infty,t} L_t^{(0)\prime} \\ &= T_t P_{\infty,t} T_t' - T_t P_{\infty,t} Z_t' K_t^{(0)\prime} \\ &= T_t P_{\infty,t} T_t', \qquad (5.16) \\ P_{*,t+1} &= T_t P_{\infty,t} L_t^{(1)\prime} + T_t P_{*,t} L_t^{(0)\prime} + R_t Q R_t' \\ &= -T_t P_{\infty,t} Z_t' K_t^{(1)\prime} + T_t P_{*,t} L_t^{(0)\prime} + R_t Q R_t' \\ &= T_t P_{*,t} L_t^{(0)\prime} + R_t Q R_t', \qquad (5.17)\end{aligned}$$

for $t = 1, \dots, d$, with $P_{\infty,1} = P_\infty = AA'$ and $P_{*,1} = P_* = R_0 Q_0 R_0'$. It might be thought that an expression of the form $F_{*,t} + \kappa^{-1} F_{**,t} + O(\kappa^{-2})$ should be used for F_t here so that two-term expansions could be carried out throughout. It can be shown however that when $M_{\infty,t} = P_{\infty,t} Z_t' = 0$, so that $F_{\infty,t} = 0$, the contribution of the term $\kappa^{-1} F_{**,t}$ is zero; we have therefore omitted it to simplify the presentation.

5.2.2 Transition to the usual Kalman filter

We now show that for nondegenerate models there is a value of d of t such that $P_{\infty,t} \neq 0$ for $t \leq d$ and $P_{\infty,t} = 0$ for $t > d$. From (5.2) the vector of diffuse elements of α_1 is δ and its dimensionality is q. For finite κ the logdensity of δ is

$$\log p(\delta) = -\frac{q}{2}\log 2\pi - \frac{q}{2}\log\kappa - \frac{1}{2\kappa}\delta'\delta,$$

since $\mathrm{E}(\delta) = 0$ and $\mathrm{Var}(\delta) = \kappa I_q$. Now consider the joint density of δ and Y_t. In an obvious notation the log conditional density of δ given Y_t is

$$\log p(\delta|Y_t) = \log p(\delta, Y_t) - \log p(Y_t),$$

for $t = 1, \dots, n$. Differentiating with respect to δ, letting $\kappa \to \infty$, equating to zero and solving for δ, gives a solution $\delta = \tilde{\delta}$ which is the conditional mode, and hence the conditional mean, of δ given Y_t.

Since $p(\delta, Y_t)$ is Gaussian, $\log p(\delta, Y_t)$ is quadratic in δ so its second derivative does not depend on δ. The reciprocal of minus the second derivative is the variance matrix of δ given Y_t. Let d be the first value of t for which this variance matrix exists. In practical cases d will usually be small relative to n. If there is no value of t for which the variance matrix exists we say that the model is degenerate, since a series of observations which does not even contain enough information to estimate the initial conditions is clearly useless.

By repeated substitution from the state equation $\alpha_{t+1} = T_t\alpha_t + R_t\eta_t$ we can express α_{t+1} as a linear function of α_1 and $\eta_1, \ldots, \eta_t$. Elements of α_1 other than those in δ and also elements of $\eta_1, \ldots, \eta_t$ have finite unconditional variances and hence have finite conditional variances given Y_t. We have also ensured that elements of δ have finite conditional variances given Y_t for $t \geq d$ by definition of d. It follows that $\text{Var}(\alpha_{t+1}|Y_t) = P_{t+1}$ is finite and hence $P_{\infty,t+1} = 0$ for $t \geq d$. On the other hand, for $t < d$, elements of $\text{Var}(\delta|Y_t)$ become infinite as $\kappa \to \infty$ from which it follows that elements of $\text{Var}(\alpha_{t+1}|Y_t)$ become infinite, so $P_{\infty,t+1} \neq 0$ for $t < d$. This establishes that for nondegenerate models there is a value d of t such that $P_{\infty,t} \neq 0$ for $t \leq d$ and $P_{\infty,t} = 0$ for $t > d$. Thus when $t > d$ we have $P_t = P_{*,t} + O(\kappa^{-1})$ so on letting $\kappa \to \infty$ we can use the usual Kalman filter (4.24) starting with $a_{d+1} = a_{d+1}^{(0)}$ and $P_{d+1} = P_{*,d+1}$. A similar discussion of this point is given by de Jong (1991).

5.2.3 A convenient representation

The updating equations for $P_{*,t+1}$ and $P_{\infty,t+1}$, for $t = 1, \ldots, d$, can be combined to obtain a very convenient representation. Let

$$P_t^\dagger = [P_{*,t} \quad P_{\infty,t}], \qquad L_t^\dagger = \begin{bmatrix} L_t^{(0)} & L_t^{(1)} \\ 0 & L_t^{(0)} \end{bmatrix}. \tag{5.18}$$

From (5.14), the limiting initial state filtering equations as $\kappa \to \infty$ can be written as

$$P_{t+1}^\dagger = T_t P_t^\dagger L_t^{\dagger\prime} + [R_t Q_t R_t' \quad 0], \qquad t = 1, \ldots, d, \tag{5.19}$$

with the initialisation $P_1^\dagger = P^\dagger = [P_* \quad P_\infty]$. For the case $F_{\infty,t} = 0$, the equations in (5.19) with the definitions in (5.18) are still valid but with

$$K_t^{(0)} = T_t M_{*,t} F_{*,t}^{-1}, \qquad L_t^{(0)} = T_t - K_t^{(0)} Z_t, \qquad L_t^{(1)} = 0.$$

This follows directly from the argument used to derive (5.15), (5.16) and (5.17). The recursion (5.19) for diffuse state filtering is due to Koopman and Durbin (2003). It is similar in form to the standard Kalman filtering (4.24) recursion, leading to simplifications in implementing the computations.

5.3 Exact initial state smoothing

5.3.1 Smoothed mean of state vector

To obtain the limiting recursions for the smoothing equation $\hat{\alpha}_t = a_t + P_t r_{t-1}$ given in (4.39) for $t = d, \ldots, 1$, we return to the recursion (4.38) for r_{t-1}, that is,

$$r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t, \qquad t = n, \dots, 1,$$

with $r_n = 0$. Since r_{t-1} depends on F_t^{-1} and L_t which can both be expressed as power series in κ^{-1} we write

$$r_{t-1} = r_{t-1}^{(0)} + \kappa^{-1} r_{t-1}^{(1)} + O(\kappa^{-2}), \qquad t = d, \dots, 1. \tag{5.20}$$

Substituting the relevant expansions into the recursion for r_{t-1} we have for the case $F_{\infty,t}$ nonsingular,

$$\begin{aligned} r_{t-1}^{(0)} + \kappa^{-1} r_{t-1}^{(1)} + \cdots = {} & Z_t' \big(\kappa^{-1} F_t^{(1)} + \kappa^{-2} F_t^{(2)} + \cdots\big)\big(v_t^{(0)} + \kappa^{-1} v_t^{(1)} + \cdots\big) \\ & + \big(L_t^{(0)} + \kappa^{-1} L_t^{(1)} + \cdots\big)'\big(r_t^{(0)} + \kappa^{-1} r_t^{(1)} + \cdots\big), \end{aligned}$$

leading to recursions for $r_t^{(0)}$ and $r_t^{(1)}$,

$$\begin{aligned} r_{t-1}^{(0)} &= L_t^{(0)\prime} r_t^{(0)}, \\ r_{t-1}^{(1)} &= Z_t' F_t^{(1)} v_t^{(0)} + L_t^{(0)\prime} r_t^{(1)} + L_t^{(1)\prime} r_t^{(0)}, \end{aligned} \tag{5.21}$$

for $t = d, \dots, 1$ with $r_d^{(0)} = r_d$ and $r_d^{(1)} = 0$.

The smoothed state vector is

$$\begin{aligned} \hat{\alpha}_t &= a_t + P_t r_{t-1} \\ &= a_t + [\kappa P_{\infty,t} + P_{*,t} + O(\kappa^{-1})]\big[r_{t-1}^{(0)} + \kappa^{-1} r_{t-1}^{(1)} + O(\kappa^{-2})\big] \\ &= a_t + \kappa P_{\infty,t}\big(r_{t-1}^{(0)} + \kappa^{-1} r_{t-1}^{(1)}\big) + P_{*,t}\big(r_{t-1}^{(0)} + \kappa^{-1} r_{t-1}^{(1)}\big) + O(\kappa^{-1}) \\ &= a_t + \kappa P_{\infty,t} r_{t-1}^{(0)} + P_{*,t} r_{t-1}^{(0)} + P_{\infty,t} r_{t-1}^{(1)} + O(\kappa^{-1}), \end{aligned} \tag{5.22}$$

where $a_t = a_t^{(0)} + \kappa^{-1} a_t^{(1)} + \cdots$. It is immediately obvious that for this expression to make sense we must have $P_{\infty,t} r_{t-1}^{(0)} = 0$ for all t. This will be the case if we can show that $\mathrm{Var}(\alpha_t | Y_n)$ is finite for all t as $\kappa \to \infty$. Analogously to the argument in Subsection 5.2.2 we can express α_t as a linear function of $\delta, \eta_0, \eta_1, \dots, \eta_{t-1}$. But $\mathrm{Var}(\delta | Y_d)$ is finite by definition of d so $\mathrm{Var}(\delta | Y_n)$ must be finite as $\kappa \to \infty$ since $d < n$. Also, $Q_j = \mathrm{Var}(\eta_j)$ is finite so $\mathrm{Var}(\eta_j | Y_n)$ is finite for $j = 0, \dots, t-1$. It follows that $\mathrm{Var}(\alpha_t | Y_n)$ is finite for all t as $\kappa \to \infty$ so from (5.22) $P_{\infty,t} r_{t-1}^{(0)} = 0$.

Letting $\kappa \to \infty$ we obtain

$$\hat{\alpha}_t = a_t^{(0)} + P_{*,t} r_{t-1}^{(0)} + P_{\infty,t} r_{t-1}^{(1)}, \qquad t = d, \dots, 1, \tag{5.23}$$

with $r_d^{(0)} = r_d$ and $r_d^{(1)} = 0$. The equations (5.21) and (5.23) can be re-formulated to obtain

$$r_{t-1}^{\dagger} = \begin{pmatrix} 0 \\ Z_t' F_t^{(1)} v_t^{(0)} \end{pmatrix} + L_t^{\dagger\prime} r_t^{\dagger}, \qquad \hat{\alpha}_t = a_t^{(0)} + P_t^{\dagger} r_{t-1}^{\dagger}, \qquad t = d, \dots, 1, \tag{5.24}$$

where

$$r_{t-1}^{\dagger} = \begin{pmatrix} r_{t-1}^{(0)} \\ r_{t-1}^{(1)} \end{pmatrix}, \quad \text{with} \quad r_d^{\dagger} = \begin{pmatrix} r_d \\ 0 \end{pmatrix},$$

and the partioned matrices $P_t^{\dagger}$ and $L_t^{\dagger}$ are defined in (5.18). This formulation is convenient since it has the same form as the standard smoothing recursion (4.39). Considering the complexity introduced into the model by the presence of the diffuse elements in α_1, it is very interesting that the state smoothing equations in (5.24) have the same basic structure as the corresponding equations (4.39). This is a useful property in constructing software for implementation of the algorithms.

To avoid extending the treatment further, and since the case $F_{\infty,t} = 0$ is rare in practice, when $P_{\infty,t} \neq 0$, we omit consideration of it here and in Subsection 5.3.2 and refer the reader to the discussion in Koopman and Durbin (2003).

5.3.2 Smoothed variance of state vector

We now consider the evaluation of the variance matrix of the estimation error $\hat{\alpha}_t - \alpha_t$ for $t = d, \dots, 1$ in the diffuse case. We shall not derive the cross-covariances between the estimation errors at different time points in the diffuse case because in practice there is little interest in these quantities.

From Subsection 4.4.3, the error variance matrix of the smoothed state vector is given by $V_t = P_t - P_t N_{t-1} P_t$ with the recursion $N_{t-1} = Z_t' F_t^{-1} Z_t + L_t' N_t L_t$, for $t = n, \dots, 1$, and $N_n = 0$. To obtain exact finite expressions for V_t and N_{t-1} where $F_{\infty,t}$ is nonsingular and $\kappa \to \infty$, for $t = d, \dots, 1$, we find that we need to take three-term expansions instead of the two-term expressions previously employed. Thus we write

$$N_t = N_t^{(0)} + \kappa^{-1} N_t^{(1)} + \kappa^{-2} N_t^{(2)} + O(\kappa^{-3}). \tag{5.25}$$

Ignoring residual terms and on substituting in the expression for N_{t-1}, we obtain the recursion for N_{t-1} as

$$\begin{aligned} N_{t-1}^{(0)} &+ \kappa^{-1} N_{t-1}^{(1)} + \kappa^{-2} N_{t-1}^{(2)} + \cdots \\ &= Z_t' \big(\kappa^{-1} F_t^{(1)} + \kappa^{-2} F_t^{(2)} + \cdots\big) Z_t + \big(L_t^{(0)} + \kappa^{-1} L_t^{(1)} + \kappa^{-2} L_t^{(2)} + \cdots\big)' \\ &\quad \times \big(N_t^{(0)} + \kappa^{-1} N_t^{(1)} + \kappa^{-2} N_t^{(2)} + \cdots\big) \big(L_t^{(0)} + \kappa^{-1} L_t^{(1)} + \kappa^{-2} L_t^{(2)} + \cdots\big), \end{aligned}$$

which leads to the set of recursions

$$
\begin{aligned}
N_{t-1}^{(0)} &= L_t^{(0)\prime} N_t^{(0)} L_t^{(0)}, \\
N_{t-1}^{(1)} &= Z_t' F_t^{(1)} Z_t + L_t^{(0)\prime} N_t^{(1)} L_t^{(0)} + L_t^{(1)\prime} N_t^{(0)} L_t^{(0)} + L_t^{(0)\prime} N_t^{(0)} L_t^{(1)}, \\
N_{t-1}^{(2)} &= Z_t' F_t^{(2)} Z_t + L_t^{(0)\prime} N_t^{(2)} L_t^{(0)} + L_t^{(0)\prime} N_t^{(1)} L_t^{(1)} + L_t^{(1)\prime} N_t^{(1)} L_t^{(0)} \\
&\qquad + L_t^{(0)\prime} N_t^{(0)} L_t^{(2)} + L_t^{(2)\prime} N_t^{(0)} L_t^{(0)} + L_t^{(1)\prime} N_t^{(0)} L_t^{(1)},
\end{aligned}
\tag{5.26}
$$

with $N_d^{(0)} = N_d$ and $N_d^{(1)} = N_d^{(2)} = 0$.

Substituting the power series in κ^{-1}, κ^{-2}, etc. and the expression $P_t = \kappa P_{\infty,t} + P_{*,t}$ into the relation $V_t = P_t - P_t N_{t-1} P_t$ we obtain

$$
\begin{aligned}
V_t &= \kappa P_{\infty,t} + P_{*,t} \\
&\quad - (\kappa P_{\infty,t} + P_{*,t})\big(N_{t-1}^{(0)} + \kappa^{-1} N_{t-1}^{(1)} + \kappa^{-2} N_{t-1}^{(2)} + \cdots\big)(\kappa P_{\infty,t} + P_{*,t}) \\
&= -\kappa^2 P_{\infty,t} N_{t-1}^{(0)} P_{\infty,t} \\
&\quad + \kappa\big(P_{\infty,t} - P_{\infty,t} N_{t-1}^{(0)} P_{*,t} - P_{*,t} N_{t-1}^{(0)} P_{\infty,t} - P_{\infty,t} N_{t-1}^{(1)} P_{\infty,t}\big) \\
&\quad + P_{*,t} - P_{*,t} N_{t-1}^{(0)} P_{*,t} - P_{*,t} N_{t-1}^{(1)} P_{\infty,t} - P_{\infty,t} N_{t-1}^{(1)} P_{*,t} \\
&\quad - P_{\infty,t} N_{t-1}^{(2)} P_{\infty,t} + O(\kappa^{-1}).
\end{aligned}
\tag{5.27}
$$

It was shown in the previous section that $V_t = \mathrm{Var}(\alpha_t | Y_n)$ is finite for $t = 1, \ldots, n$. Thus the two matrix terms associated with κ and κ^2 in (5.27) must be zero. Letting $\kappa \to \infty$, the smoothed state variance matrix is given by

$$
V_t = P_{*,t} - P_{*,t} N_{t-1}^{(0)} P_{*,t} - P_{*,t} N_{t-1}^{(1)} P_{\infty,t} - P_{\infty,t} N_{t-1}^{(1)} P_{*,t} - P_{\infty,t} N_{t-1}^{(2)} P_{\infty,t}.
\tag{5.28}
$$

Koopman and Durbin (2003) have shown by exploiting the equality $P_{\infty,t} L_t^{(0)} N_t^{(0)} = 0$, that when the recursions for $N_t^{(1)}$ and $N_t^{(2)}$ in (5.26) are used to calculate the terms in $N_{t-1}^{(1)}$ and $N_{t-1}^{(2)}$, respectively, in (5.28), various items vanish and that the effect is that we can proceed in effect as if the recursions are

$$
\begin{aligned}
N_{t-1}^{(0)} &= L_t^{(0)\prime} N_t^{(0)} L_t^{(0)}, \\
N_{t-1}^{(1)} &= Z_t' F_t^{(1)} Z_t + L_t^{(0)\prime} N_t^{(1)} L_t^{(0)} + L_t^{(1)\prime} N_t^{(0)} L_t^{(0)}, \\
N_{t-1}^{(2)} &= Z_t' F_t^{(2)} Z_t + L_t^{(0)\prime} N_t^{(2)} L_t^{(0)} + L_t^{(0)\prime} N_t^{(1)} L_t^{(1)} + L_t^{(1)\prime} N_t^{(1)\prime} L_t^{(0)} + L_t^{(1)\prime} N_t^{(0)} L_t^{(1)},
\end{aligned}
\tag{5.29}
$$

and that we can compute V_t by

$$
V_t = P_{*,t} - P_{*,t} N_{t-1}^{(0)} P_{*,t} - \big(P_{\infty,t} N_{t-1}^{(1)} P_{*,t}\big)' - P_{\infty,t} N_{t-1}^{(1)} P_{*,t} - P_{\infty,t} N_{t-1}^{(2)} P_{\infty,t}.
\tag{5.30}
$$

Thus the matrices calculated from (5.26) can be replaced by the ones computed by (5.29) to obtain the correct value for V_t. The new recursions in (5.29) are convenient since the matrix $L_t^{(2)}$ drops out from our calculations for $N_t^{(2)}$.

Indeed the matrix recursions for $N_t^{(1)}$ and $N_t^{(2)}$ in (5.29) are not the same as the recursions for $N_t^{(1)}$ and $N_t^{(2)}$ in (5.26). Also, it can be noticed that matrix $N_t^{(1)}$ in (5.26) is symmetric while $N_t^{(1)}$ in (5.29) is not symmetric. However, the same notation is employed because $N_t^{(1)}$ is only relevant for computing V_t and matrix $P_{\infty,t}N_{t-1}^{(1)}$ is the same when $N_{t-1}^{(1)}$ is computed by (5.26) as when it is computed by (5.29). The same argument applies to matrix $N_t^{(2)}$.

It can now be easily verified that equations (5.30) and the modified recursions (5.29) can be reformulated as

$$N_{t-1}^{\dagger} = \begin{bmatrix} 0 & Z_t'F_t^{(1)}Z_t \\ Z_t'F_t^{(1)}Z_t & Z_t'F_t^{(2)}Z_t \end{bmatrix} + L_t^{\dagger\prime}N_t^{\dagger}L_t^{\dagger}, \qquad V_t = P_{*,t} - P_t^{\dagger}N_{t-1}^{\dagger}P_t^{\dagger\prime}, \tag{5.31}$$

for $t = d, \dots, 1$, where

$$N_{t-1}^{\dagger} = \begin{bmatrix} N_{t-1}^{(0)} & N_{t-1}^{(1)\prime} \\ N_{t-1}^{(1)} & N_{t-1}^{(2)} \end{bmatrix}, \quad \text{with} \quad N_d^{\dagger} = \begin{bmatrix} N_d & 0 \\ 0 & 0 \end{bmatrix},$$

and the partioned matrices $P_t^{\dagger}$ and $L_t^{\dagger}$ are defined in (5.18). Again, this formulation has the same form as the standard smoothing recursion (4.43) which is a useful property when writing software. The formulations (5.24) and (5.31) are given by Koopman and Durbin (2003).

5.4 Exact initial disturbance smoothing

Calculating the smoothed disturbances does not require as much computing as calculating the smoothed state vector when the initial state vector is diffuse. This is because the smoothed disturbance equations do not involve matrix multiplications which depend on terms in κ or higher order terms. From (4.58) the smoothed estimator is $\hat{\varepsilon}_t = H_t(F_t^{-1}v_t - K_t'r_t)$ where $F_t^{-1} = O(\kappa^{-1})$ for $F_{\infty,t}$ positive definite, $F_t^{-1} = F_{*,t}^{-1} + O(\kappa^{-1})$ for $F_{\infty,t} = 0$, $K_t = K_t^{(0)} + O(\kappa^{-1})$ and $r_t = r_t^{(0)} + O(\kappa^{-1})$ so, as $\kappa \to \infty$, we have

$$\hat{\varepsilon}_t = \begin{cases} -H_tK_t^{(0)\prime}r_t^{(0)} & \text{if } F_{\infty,t} \text{ is nonsingular,} \\ H_t\big(F_{*,t}^{-1}v_t - K_t^{(0)\prime}r_t^{(0)}\big) & \text{if } F_{\infty,t} = 0, \end{cases}$$

for $t = d, \dots, 1$. Other results for disturbance smoothing are obtained in a similar way and for convenience we collect them together in the form

$$\begin{aligned}
\hat{\varepsilon}_t &= -H_t K_t^{(0)\prime} r_t^{(0)}, \\
\hat{\eta}_t &= Q_t R_t' r_t^{(0)}, \\
\mathrm{Var}(\varepsilon_t | Y_n) &= H_t - H_t K_t^{(0)\prime} N_t^{(0)} K_t^{(0)} H_t, \\
\mathrm{Var}(\eta_t | Y_n) &= Q_t - Q_t R_t' N_t^{(0)} R_t Q_t,
\end{aligned}$$

for the case where $F_{\infty,t} \neq 0$ and

$$\begin{aligned}
\hat{\varepsilon}_t &= H_t \big(F_{*,t}^{-1} v_t - K_t^{(0)\prime} r_t^{(0)} \big), \\
\hat{\eta}_t &= Q_t R_t' r_t^{(0)}, \\
\mathrm{Var}(\varepsilon_t | Y_n) &= H_t - H_t \big(F_{*,t}^{-1} + K_t^{(0)\prime} N_t^{(0)} K_t^{(0)} \big) H_t, \\
\mathrm{Var}(\eta_t | Y_n) &= Q_t - Q_t R_t' N_t^{(0)} R_t Q_t,
\end{aligned}$$

for the case $F_{\infty,t} = 0$ and for $t = d, \dots, 1$. It is fortunate that disturbance smoothing in the diffuse case does not require as much extra computing as for state smoothing. This is particularly convenient when the score vector is computed repeatedly within the process of parameter estimation as we will discuss in Subsection 7.3.3.

5.5 Exact initial simulation smoothing

5.5.1 Modifications for diffuse initial conditions

When the initial state vector is diffuse it turns out that the simulation smoother of Section 4.9 can be used without the complexities of Section 5.3 required for diffuse state smoothing. Let us begin by looking at diffuse filtering and smoothing from an intuitive point of view. Suppose we were to initialise model (4.12) with an entirely arbitrary value $\alpha_1 = \alpha_1^*$ and then apply formulae in Sections 5.2 to 5.4 to obtain diffuse filtered and smoothed values for the resulting observational vector Y_n. It seems evident intuitively that the filtered and smoothed values that emerge cannot depend on the value of α_1^* that we have chosen.

These ideas suggest the following conjecture for the exact treatment of simulation smoothing in the diffuse case. Set the diffuse elements of α_1 equal to arbitrary values, say zeros, and use the diffuse filters and smoothers developed in Sections 5.2 to 5.4 for the calculation of $\widetilde{w}$ and $\widetilde{\alpha}$ by the methods of Section 4.9, respectively; then these are the exact values required. The validity of this conjecture is proved in Appendix 2 of Durbin and Koopman (2002); details are intricate so they will not be repeated here.

5.5.2 Exact initial simulation smoothing

We first consider how to obtain a simulated sample of α given Y_n using the method of de Jong–Shephard. Taking $\alpha = (\alpha_1, \dots, \alpha_n, \alpha_{n+1})'$ as before, define $\alpha_{/1}$ as α but without α_1. It follows that

$$p(\alpha|Y_n) = p(\alpha_1|Y_n)p(\alpha_{/1}|Y_n, \alpha_1). \tag{5.32}$$

Obtain a simulated value $\tilde{\alpha}_1$ of α_1 by drawing a sample value from $p(\alpha_1|Y_n) = \mathrm{N}(\hat{\alpha}_1, V_1)$ for which $\hat{\alpha}_1$ and V_1 are computed by the exact initial state smoother as developed in Section 5.3. Next initialise the Kalman filter with $a_1 = \tilde{\alpha}_1$ and $P_1 = 0$, since we now treat $\tilde{\alpha}_1$ as given, and apply the Kalman filter and simulation smoother as decribed in Section 4.9. This procedure for obtaining a sample value of α given Y_n is justified by equation (5.32). To obtain multiple samples, we repeat this procedure. This requires computing a new value of $\tilde{\alpha}_1$, and new values of v_t from the Kalman filter for each new draw. The Kalman filter quantities F_t, K_t and P_{t+1} do not need to be recomputed.

A similar procedure can be followed for simulating disturbance vectors given Y_n: we initialise the Kalman filter with $a_1 = \tilde{\alpha}_1$ and $P_1 = 0$ as above and then use the simulation smoothing recursions of Section 4.9 to generate samples for the disturbances.

5.6 Examples of initial conditions for some models

In this section we give some examples of the exact initial Kalman filter for $t = 1, \dots, d$ for a range of state space models.

5.6.1 Structural time series models

Structural time series models are usually set up in terms of nonstationary components. Therefore, most of the models in this class have the initial state vector equals δ, that is, $\alpha_1 = \delta$ so that $a_1 = 0$, $P_* = 0$ and $P_\infty = I_m$. We then proceed with the algorithms provided by Sections 5.2, 5.3 and 5.4.

To illustrate the exact initial Kalman filter in detail we consider the local linear trend model (3.2) with system matrices

$$Z_t = (1 \quad 0), \qquad T_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad Q_t = \sigma_\varepsilon^2 \begin{bmatrix} q_\xi & 0 \\ 0 & q_\zeta \end{bmatrix},$$

and with $H_t = \sigma_\varepsilon^2$, $R_t = I_2$, where $q_\xi = \sigma_\xi^2/\sigma_\varepsilon^2$ and $q_\zeta = \sigma_\zeta^2/\sigma_\varepsilon^2$. The exact initial Kalman filter is started with

$$a_1 = 0, \qquad P_{*,1} = 0, \qquad P_{\infty,1} = I_2,$$

and the first update is based on

$$K_1^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad L_1^{(0)} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad K_1^{(1)} = -\sigma_\varepsilon^2 \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$
$$L_1^{(1)} = \sigma_\varepsilon^2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

such that

$$a_2 = \begin{pmatrix} y_1 \\ 0 \end{pmatrix}, \qquad P_{*,2} = \sigma_\varepsilon^2 \begin{bmatrix} 1+q_\xi & 0 \\ 0 & q_\zeta \end{bmatrix}, \qquad P_{\infty,2} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

The second update gives the quantities

$$K_2^{(0)} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad L_2^{(0)} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix},$$

and

$$K_2^{(1)} = -\sigma_\varepsilon^2 \begin{pmatrix} 3+q_\xi \\ 2+q_\xi \end{pmatrix}, \qquad L_2^{(1)} = \sigma_\varepsilon^2 \begin{bmatrix} 3+q_\xi & 0 \\ 2+q_\xi & 0 \end{bmatrix},$$

together with the state update results

$$a_3 = \begin{pmatrix} 2y_2 - y_1 \\ y_2 - y_1 \end{pmatrix}, \qquad P_{*,3} = \sigma_\varepsilon^2 \begin{bmatrix} 5+2q_\xi+q_\zeta & 3+q_\xi+q_\zeta \\ 3+q_\xi+q_\zeta & 2+q_\xi+2q_\zeta \end{bmatrix},$$
$$P_{\infty,3} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

It follows that the usual Kalman filter (4.24) can be used for $t = 3, \dots, n$.

5.6.2 Stationary ARMA models

The univariate stationary ARMA model with zero mean of order p and q is given by

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \zeta_t + \theta_1 \zeta_{t-1} + \cdots + \theta_q \zeta_{t-q}, \quad \zeta_t \sim \mathrm{N}(0, \sigma^2).$$

The state space form is

$$y_t = (1, 0, \dots, 0)\alpha_t,$$
$$\alpha_{t+1} = T\alpha_t + R\zeta_{t+1},$$

where the system matrices T and R are given by (3.20) with $r = \max(p, q+1)$. All elements of the state vector are stationary so that the part $a + A\delta$ in (5.2) is zero and $R_0 = I_m$. The unconditional distribution of the initial state vector α_1 is therefore given by

$$\alpha_1 \sim \mathrm{N}(0, \sigma^2 Q_0),$$

where, since $\mathrm{Var}(\alpha_{t+1}) = \mathrm{Var}(T\alpha_t + R\zeta_{t+1})$, $\sigma^2 Q_0 = \sigma^2 T Q_0 T' + \sigma^2 RR'$. This equation needs to be solved for Q_0. It can be shown that a solution can be obtained by solving the linear equation $(I_{m^2} - T \otimes T)\,\mathrm{vec}(Q_0) = \mathrm{vec}(RR')$ for Q_0, where $\mathrm{vec}(Q_0)$ and $\mathrm{vec}(RR')$ are the stacked columns of Q_0 and RR' and where

$$T \otimes T = \begin{bmatrix} t_{11}T & \dots & t_{1m}T \\ t_{21}T & \dots & t_{2m}T \\ \vdots & & \\ t_{m1}T & \dots & t_{mm}T \end{bmatrix},$$

with t_{ij} denoting the (i,j) element of matrix T; see, for example, Magnus and Neudecker (1988) who give a general treatment of problems of this type. The Kalman filter is initialised by $a_1 = 0$ and $P_1 = Q_0$.

As an example, consider the ARMA$(1,1)$ model

$$y_t = \phi y_{t-1} + \zeta_t + \theta\zeta_{t-1}, \qquad \zeta_t \sim \mathrm{N}(0, \sigma^2).$$

Then

$$T = \begin{bmatrix} \phi & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad R = \begin{pmatrix} 1 \\ \theta \end{pmatrix},$$

so the solution is

$$Q_0 = \begin{bmatrix} (1-\phi^2)^{-1}(1+\theta^2+2\phi\theta) & \theta \\ \theta & \theta^2 \end{bmatrix}.$$

5.6.3 Nonstationary ARIMA models

The univariate nonstationary ARIMA model of order p, d and q can be put in the form

$$y_t^* = \phi_1 y_{t-1}^* + \dots + \phi_p y_{t-p}^* + \zeta_t + \theta_1\zeta_{t-1} + \dots + \theta_q\zeta_{t-q}, \qquad \zeta_t \sim \mathrm{N}(0, \sigma^2).$$

where $y_t^* = \Delta^d y_t$. The state space form of the ARIMA model with $p = 2$, $d = 1$ and $q = 1$ is given in Section 3.4 with the state vector given by

$$\alpha_t = \begin{pmatrix} y_{t-1} \\ y_t^* \\ \phi_2 y_{t-1}^* + \theta_1\zeta_t \end{pmatrix},$$

where $y_t^* = \Delta y_t = y_t - y_{t-1}$. The first element of the initial state vector α_1, that is y_0, is nonstationary while the other elements are stationary. Therefore, the initial vector $\alpha_1 = a + A\delta + R_0\eta_0$ is given by

$$\alpha_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \delta + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \eta_0, \qquad \eta_0 \sim \mathrm{N}(0, Q_0),$$

where Q_0 is the 2×2 unconditional variance matrix for an ARMA model with $p = 2$ and $q = 1$ which we obtain from Subsection 5.6.2. When δ is diffuse, the mean vector and variance matrix are

$$a_1 = 0, \qquad P_1 = \kappa P_\infty + P_*,$$

where

$$P_\infty = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad P_* = \begin{bmatrix} 0 & 0 \\ 0 & Q_0 \end{bmatrix}.$$

Analysis then proceeds using the exact initial Kalman filter and smoother of Sections 5.2, 5.3 and 5.4. The initial state specification for ARIMA models with $d = 1$ but with other values for p and q is obtained in a similar way.

From Section 3.4, the initial state vector for the ARIMA model with $p = 2$, $d = 2$ and $q = 1$ is given by

$$\alpha_1 = \begin{pmatrix} y_0 \\ \Delta y_0 \\ y_1^* \\ \phi_2 y_0^* + \theta_1 \zeta_1 \end{pmatrix}.$$

The first two elements of α_1, that is, y_0 and Δy_0, are nonstationary and we therefore treat them as diffuse. Thus we write

$$\alpha_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \delta + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \eta_0, \qquad \eta_0 \sim \mathrm{N}(0, Q_0),$$

where Q_0 is as for the previous case. It follows that the mean vector and variance matrix of α_1 are

$$a_1 = 0, \qquad P_1 = \kappa P_\infty + P_*,$$

where

$$P_\infty = \begin{bmatrix} I_2 & 0 \\ 0 & 0 \end{bmatrix}, \qquad P_* = \begin{bmatrix} 0 & 0 \\ 0 & Q_0 \end{bmatrix}.$$

We then proceed with the methods of Sections 5.2, 5.3 and 5.4. The initial conditions for non-seasonal ARIMA models with other values for p, d and q and seasonal models are derived in similar ways.

5.6.4 Regression model with ARMA errors

The regression model with k explanatory variables and ARMA(p, q) errors (3.30) can be written in state space form as in Subsection 3.6.2. The initial state vector is

$$\alpha_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{bmatrix} I_k \\ 0 \end{bmatrix} \delta + \begin{bmatrix} 0 \\ I_r \end{bmatrix} \eta_0, \qquad \eta_0 \sim \mathrm{N}(0, Q_0),$$

where Q_0 is obtained as in Subsection 5.6.2 and $r = \max(p, q+1)$. When δ is treated as diffuse we have $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ where $a_1 = 0$ and $P_1 = \kappa P_\infty + P_*$ with

$$P_\infty = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}, \qquad P_* = \begin{bmatrix} 0 & 0 \\ 0 & Q_0 \end{bmatrix}.$$

We then proceed as described in the last section.

To illustrate the use of the exact initial Kalman filter we consider the simple case of an AR(1) model with a constant, that is

$$\begin{aligned} y_t &= \mu + \xi_t, \\ \xi_t &= \phi \xi_{t-1} + \zeta_t, \qquad \zeta_t \sim \mathrm{N}(0, \sigma^2). \end{aligned}$$

In state space form we have

$$\alpha_t = \begin{pmatrix} \mu \\ \xi_t \end{pmatrix}$$

and the system matrices are given by

$$Z_t = (1 \quad 1), \qquad T_t = \begin{bmatrix} 1 & 0 \\ 0 & \phi \end{bmatrix}, \qquad R_t = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

with $H_t = 0$ and $Q_t = \sigma^2$. The exact initial Kalman filter is started with

$$a_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad P_{*,1} = c \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \qquad P_{\infty,1} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

where $c = \sigma^2/(1 - \phi^2)$. The first update is based on

$$K_1^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad L_1^{(0)} = \begin{bmatrix} 0 & -1 \\ 0 & \phi \end{bmatrix}, \qquad K_1^{(1)} = c \begin{pmatrix} -1 \\ \phi \end{pmatrix},$$
$$L_1^{(1)} = c \begin{bmatrix} 1 & 1 \\ -\phi & -\phi \end{bmatrix},$$

such that

$$a_2 = \begin{pmatrix} y_1 \\ 0 \end{pmatrix}, \qquad P_{*,2} = \frac{\sigma^2}{1 - \phi^2} \begin{bmatrix} 1 & -\phi \\ -\phi & 1 \end{bmatrix}, \qquad P_{\infty,2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

It follows that the usual Kalman filter (4.24) can be used for $t = 2, \ldots, n$.

5.6.5 Spline smoothing

The initial state vector for the spline model (3.44) is simply $\alpha_1 = \delta$, implying that $a_1 = 0$, $P_* = 0$ and $P_\infty = I_2$.

5.7 Augmented Kalman filter and smoother

5.7.1 Introduction

An alternative approach for dealing with the initialisation problem is owed to Rosenberg (1973), de Jong (1988b) and de Jong (1991). As in (5.2), the initial state vector is defined as

$$\alpha_1 = a + A\delta + R_0\eta_0, \qquad \eta_0 \sim \mathrm{N}(0, Q_0). \tag{5.33}$$

Rosenberg (1973) treats δ as a fixed unknown vector and he employs maximum likelihood to estimate δ while de Jong (1991) treats δ as a diffuse vector. Since the treatments of Rosenberg and de Jong are both based on the idea of augmenting the observed vector, we will refer to their procedures collectively as the *augmentation approach*. The approach of Rosenberg offers relief to analysts who feel uncomfortable about using diffuse initialising densities on the ground that infinite variances have no counterpart in real data. In fact, as we shall show, the two approaches give effectively the same answer so these analysts could regard the diffuse assumption as a device for achieving initialisation based on maximum likelihood estimates of the unknown initial state elements. The results are developed from a classical point of view for the case where the observations are normally distributed. Corresponding results for MVLUE and Bayesian analyses can be obtained by applying Lemmas 2, 3 and 4 of Section 4.2.

5.7.2 Augmented Kalman filter

In this subsection we establish the groundwork for both the Rosenberg and the de Jong techniques. For given δ, apply the Kalman filter, (4.24) with $a_1 = \mathrm{E}(\alpha_1) = a + A\delta$, $P_1 = \mathrm{Var}(\alpha_1) = P_* = R_0Q_0R_0'$ and denote the resulting value of a_t from the filter output by $a_{\delta,t}$. Since $a_{\delta,t}$ is a linear function of the observations and $a_1 = a + A\delta$ we can write

$$a_{\delta,t} = a_{a,t} + A_{A,t}\delta, \tag{5.34}$$

where $a_{a,t}$ is the value of a_t in the filter output obtained by taking $a_1 = a$, $P_1 = P_*$ and where the jth column of $A_{A,t}$ is the value of a_t in the filter output obtained from an observational vector of zeros with $a_1 = A_j$, $P_1 = P_*$, where A_j is the jth column of A. Denote the value of v_t in the filter output obtained by taking $a_1 = a + A\delta$, $P_1 = P_*$ by $v_{\delta,t}$. Analogously to (5.34) we can write

$$v_{\delta,t} = v_{a,t} + V_{A,t}\delta, \tag{5.35}$$

where $v_{a,t}$ and $V_{A,t}$ are given by the same Kalman filters that produced $a_{a,t}$ and $A_{A,t}$.

The matrices $(a_{a,t}, A_{A,t})$ and $(v_{a,t}, V_{A,t})$ can be computed in one pass through a Kalman filter which inputs the observation vector y_t augmented by zero values. This is possible because for each Kalman filter producing a particular column of $(a_{a,t}, A_{A,t})$, the same variance initialisation $P_1 = P_*$ applies, so the variance output, which we denote by $F_{\delta,t}$, $K_{\delta,t}$ and $P_{\delta,t+1}$, is the same for each Kalman filter. Replacing the Kalman filter equations for the vectors v_t and a_t by the corresponding equations for the matrices $(a_{a,t}, A_{A,t})$ and $(v_{a,t}, V_{A,t})$ leads to the equations

$$\begin{aligned}(v_{a,t}, V_{A,t}) &= (y_t, 0) - Z_t(a_{a,t}, A_{A,t}),\\ (a_{a,t+1}, A_{A,t+1}) &= T_t(a_{a,t}, A_{A,t}) + K_{\delta,t}(v_{a,t}, V_{A,t}),\end{aligned} \tag{5.36}$$

where $(a_{a,1}, A_{A,1}) = (a, A)$; the recursions corresponding to F_t, K_t and P_{t+1} remain as for the standard Kalman filter, that is,

$$\begin{aligned}F_{\delta,t} &= Z_t P_{\delta,t} Z_t' + H_t,\\ K_{\delta,t} &= T_t P_{\delta,t} Z_t' F_{\delta,t}^{-1}, \qquad L_{\delta,t} = T_t - K_{\delta,t} Z_t,\\ P_{\delta,t+1} &= T_t P_{\delta,t} L_{\delta,t}' + R_t Q_t R_t',\end{aligned} \tag{5.37}$$

for $t = 1, \ldots, n$ with $P_{\delta,1} = P_*$. We have included the suffix δ in these expressions not because they depend mathematically on δ but because they have been calculated on the assumption that δ is fixed. The modified Kalman filter (5.36) and (5.37) will be referred to as the *augmented Kalman filter* in this book.

5.7.3 Filtering based on the augmented Kalman filter

With these preliminaries, let us first consider the diffuse case (5.2) with $\delta \sim \mathrm{N}(0, \kappa I_q)$ where $\kappa \to \infty$; we will consider later the case where δ is fixed and is estimated by maximum likelihood. From (5.34) we obtain for given κ,

$$a_{t+1} = \mathrm{E}(\alpha_{t+1}|Y_t) = a_{a,t+1} + A_{A,t+1}\bar{\delta}_t, \tag{5.38}$$

where $\bar{\delta}_t = \mathrm{E}(\delta|Y_t)$. Now

$$\begin{aligned}\log p(\delta|Y_t) &= \log p(\delta) + \log p(Y_t|\delta) - \log p(Y_t)\\ &= \log p(\delta) + \sum_{j=1}^{t} \log p(v_{\delta,j}) - \log p(Y_t)\\ &= -\frac{1}{2\kappa}\delta'\delta - b_t'\delta - \frac{1}{2}\delta' S_{A,t}\delta + \text{terms independent of } \delta,\end{aligned} \tag{5.39}$$

where

$$b_t = \sum_{j=1}^{t} V_{A,j}' F_{\delta,j}^{-1} v_{a,j}, \qquad S_{A,t} = \sum_{j=1}^{t} V_{A,j}' F_{\delta,j}^{-1} V_{A,j}. \tag{5.40}$$

Since densities are normal, the mean of $p(\delta|Y_t)$ is equal to the mode, and this is the value of δ which maximises $\log p(\delta|Y_t)$, so on differentiating (5.39) with respect to δ and equating to zero, we have

$$\bar{\delta}_t = -\left(S_{A,t} + \frac{1}{\kappa}I_q\right)^{-1} b_t. \tag{5.41}$$

Also,

$$\begin{aligned} P_{t+1} &= \mathrm{E}[(a_{t+1} - \alpha_{t+1})(a_{t+1} - \alpha_{t+1})'] \\ &= \mathrm{E}[\{a_{\delta,t+1} - \alpha_{t+1} - A_{A,t+1}(\delta - \bar{\delta}_t)\}\{a_{\delta,t+1} - \alpha_{t+1} - A_{A,t+1}(\delta - \bar{\delta}_t)\}'] \\ &= P_{\delta,t+1} + A_{A,t+1}\mathrm{Var}(\delta|Y_t)A'_{A,t+1} \\ &= P_{\delta,t+1} + A_{A,t+1}\left(S_{A,t} + \frac{1}{\kappa}I_q\right)^{-1} A'_{A,t+1}, \end{aligned} \tag{5.42}$$

since $\mathrm{Var}(\delta|Y_t) = (S_{A,t} + \frac{1}{\kappa}I_q)^{-1}$.

Letting $\kappa \to \infty$ we have

$$\bar{\delta}_t = -S_{A,t}^{-1}b_t, \tag{5.43}$$

$$\mathrm{Var}(\delta|Y_t) = S_{A,t}^{-1}, \tag{5.44}$$

when $S_{A,t}$ is nonsingular. The calculations of b_t and $S_{A,t}$ are easily incorporated into the augmented Kalman filter (5.36) and (5.37). It follows that

$$a_{t+1} = a_{a,t+1} - A_{A,t+1}S_{A,t}^{-1}b_t, \tag{5.45}$$

$$P_{t+1} = P_{\delta,t+1} + A_{A,t+1}S_{A,t}^{-1}A'_{A,t+1}, \tag{5.46}$$

as $\kappa \to \infty$. For $t < d$, $S_{A,t}$ is singular so values of a_{t+1} and P_{t+1} given by (5.45) and (5.46) do not exist. However, when $t = d$, a_{d+1} and P_{d+1} exist and consequently when $t > d$ the values a_{t+1} and P_{t+1} can be calculated by the standard Kalman filter for $t = d+1, \dots, n$. Thus we do not need to use the augmented Kalman filter (5.36) for $t = d+1, \dots, n$. These results are due to de Jong (1991) but our derivation here is more transparent.

We now consider a variant of the maximum likelihood method for initialising the filter due to Rosenberg (1973). In this technique, δ is regarded as fixed and unknown and we employ maximum likelihood given Y_t to obtain estimate $\hat{\delta}_t$. The loglikelihood of Y_t given δ is

$$\log p(Y_t|\delta) = \sum_{j=1}^{t} \log p(v_{\delta,j}) = -b_t'\delta - \frac{1}{2}\delta' S_{A,t}\delta + \text{terms independent of } \delta,$$

which is the same as (5.39) apart from the term $-\delta'\delta/(2\kappa)$. Differentiating with respect to δ, equating to zero and taking the second derivative gives

$$\hat{\delta}_t = -S_{A,t}^{-1} b_t, \qquad \mathrm{Var}(\hat{\delta}_t) = S_{A,t}^{-1},$$

when $S_{A,t}$ is nonsingular, that is for $t = d, \ldots, n$. These values are the same as $\bar{\delta}_t$ and $\mathrm{Var}(\delta_t | Y_t)$ when $\kappa \to \infty$. In practice we choose t to be the smallest value for which $\hat{\delta}_t$ exists, which is d. It follows that the values of a_{t+1} and P_{t+1} for $t \geq d$ given by this approach are the same as those obtained in the diffuse case. Thus the solution of the initialisation problem given in Section 5.2 also applies to the case where δ is fixed and unknown. From a computational point of view the calculations of Section 5.2 are more efficient than those for the augmented device described in this section when the model is reasonably large. A comparison of the computational efficiency is given in Subsection 5.7.5. Rosenberg (1973) used a procedure which differed slightly from this. Although he employed essentially the same augmentation technique, in the notation above he estimated δ by the value $\hat{\delta}_n$ based on all the data.

5.7.4 Illustration: the local linear trend model

To illustrate the augmented Kalman filter we consider the same local linear trend model as in Subsection 5.6.1. The system matrices of the local linear trend model (3.2) are given by

$$Z = (1 \quad 0), \qquad T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad Q = \sigma_\varepsilon^2 \begin{bmatrix} q_\xi & 0 \\ 0 & q_\zeta \end{bmatrix},$$

with $H = \sigma_\varepsilon^2$ and $R = I_2$ and where $q_\xi = \sigma_\xi^2/\sigma_\varepsilon^2$ and $q_\zeta = \sigma_\zeta^2/\sigma_\varepsilon^2$. The augmented Kalman filter is started with

$$(a_{a,1}, A_{A,1}) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad P_{\delta,1} = \sigma_\varepsilon^2 \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

and the first update is based on

$$(v_{a,1}, V_{A,1}) = (y_1 \quad -1 \quad 0), \qquad F_{\delta,1} = \sigma_\varepsilon^2, \qquad K_{\delta,1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$L_{\delta,1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

so

$$b_1 = -\frac{1}{\sigma_\varepsilon^2} \begin{pmatrix} y_1 \\ 0 \end{pmatrix}, \qquad S_{A,1} = \frac{1}{\sigma_\varepsilon^2} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

and

$$(a_{a,2}, A_{A,2}) = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \qquad P_{\delta,2} = \sigma_\varepsilon^2 \begin{bmatrix} q_\xi & 0 \\ 0 & q_\zeta \end{bmatrix}.$$

The second update gives the quantities

$$(v_{a,2}, V_{A,2}) = (y_2 \quad -1 \quad -1), \qquad F_{\delta,2} = \sigma_\varepsilon^2(1+q_\xi),$$
$$K_{\delta,2} = \frac{q_\xi}{1+q_\xi}\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad L_{\delta,2} = \begin{bmatrix} \frac{1}{1+q_\xi} & 1 \\ 0 & 1 \end{bmatrix},$$

with

$$b_2 = \frac{-1}{1+q_\xi}\begin{pmatrix} (1+q_\xi)y_1 + y_2 \\ y_2 \end{pmatrix}, \qquad S_{A,2} = \frac{1}{\sigma_\varepsilon^2(1+q_\xi)}\begin{bmatrix} 2+q_\xi & 1 \\ 1 & 1 \end{bmatrix},$$

and the state update results

$$(a_{a,3}, A_{A,3}) = \frac{1}{1+q_\xi}\begin{bmatrix} q_\xi y_2 & 1 & 2+q_\xi \\ 0 & 0 & 1+q_\xi \end{bmatrix}, \quad P_{\delta,3} = \sigma_\varepsilon^2 \begin{bmatrix} q_\xi + \frac{q_\xi}{1+q_\xi} + q_\zeta & q_\zeta \\ q_\zeta & 2q_\zeta \end{bmatrix}.$$

The augmented part can be collapsed since $S_{A,2}$ is nonsingular, giving

$$S_{A,2}^{-1} = \sigma_\varepsilon^2 \begin{bmatrix} 1 & -1 \\ -1 & 2+q_\xi \end{bmatrix}, \qquad \bar{\delta}_2 = -S_{A,2}^{-1} b_2 = \begin{pmatrix} y_1 \\ y_2 - y_1 \end{pmatrix}.$$

It follows that

$$a_3 = a_{a,3} + A_{A,3}\bar{\delta}_2 = \begin{pmatrix} 2y_2 - y_1 \\ y_2 - y_1 \end{pmatrix},$$
$$P_3 = P_{\delta,3} + A_{A,3}S_{A,2}^{-1}A'_{A,3} = \sigma_\varepsilon^2 \begin{bmatrix} 5 + 2q_\xi + q_\zeta & 3 + q_\xi + q_\zeta \\ 3 + q_\xi + q_\zeta & 2 + q_\xi + 2q_\zeta \end{bmatrix}.$$

and the usual Kalman filter (4.24) can be used for $t = 3, \ldots, n$. These results are exactly the same as those obtained in Section 5.6.1, though the computations take longer as we will now show.

5.7.5 Comparisons of computational efficiency

The adjusted Kalman filters of Section 5.2 and Subsection 5.7.2 both require more computations than the Kalman filter (4.24) with known initial conditions. Of course the adjustments are only required for a limited number of updates. The additional computations for the exact initial Kalman filter are due to updating the matrix $P_{\infty,t+1}$ and computing the matrices $K_t^{(1)}$ and $L_t^{(1)}$ when $F_{\infty,t} \neq 0$, for $t = 1, \ldots, d$. For many practical state space models the system matrices Z_t and T_t are sparse selection matrices containing many zeros and ones; this is the case for the models discussed in Chapter 3. Therefore, calculations involving Z_t and T_t are particularly cheap for most models. Table 5.1 compares the number of additional multiplications (compared to the Kalman filter with known initial

Table 5.1 Number of additional multiplications for filtering.

Model	Exact initial	Augmenting	Difference (%)
Local level	3	7	57
Local linear trend	18	46	61
Basic seasonal (s = 4)	225	600	63
Basic seasonal (s = 12)	3549	9464	63

conditions) required for filtering using the devices of Section 5.2 and Subsection 5.7.2 applied to several structural time series models which are discussed in Section 3.2. The results in Table 5.1 show that the additional number of computations for the exact initial Kalman filter of Section 5.2 is less than half the extra computations required for the augmentation device of Subsection 5.7.2. Such computational efficiency gains are important when the Kalman filter is used many times as is the case for parameter estimation; a detailed discussion of estimation is given in Chapter 7. It will also be argued in Subsection 7.3.5 that many computations for the exact initial Kalman filter only need to be done once for a specific model since the computed values remain the same when the parameters of the model change. This argument does not apply to the augmentation device and this is another important reason why our approach in Section 5.2 is more efficient than the augmentation approach.

5.7.6 Smoothing based on the augmented Kalman filter

The smoothing algorithms can also be developed using the augmented approach. The smoothing recursion for r_{t-1} in (4.69) needs to be augmented in the same way as is done for v_t and a_t of the Kalman filter. When the augmented Kalman filter is applied for $t = 1, \ldots, n$, the modifications for smoothing are straightforward after computing $\hat{\delta}_n$ and $\text{Var}(\hat{\delta}_n)$ and then applying similar expressions to those of (5.45) and (5.46). The collapse of the augmented Kalman filter to the standard Kalman filter is computationally efficient for filtering but, as a result, the estimates $\hat{\delta}_n$ and $\text{Var}(\hat{\delta}_n)$ are not available for calculating the smoothed estimates of the state vector. It is not therefore straightforward to do smoothing when the collapsing device is used in the augmentation approach. A solution for this problem has been given by Chu-Chun-Lin and de Jong (1993).

6 Further computational aspects

6.1 Introduction

In this chapter we will discuss a number of remaining computational aspects of the Kalman filter and smoother. We shall continue the treatment based on classical analysis on the assumption that results for linear estimation or Bayesian analysis can be obtained by application of Lemmas 2, 3 and 4. Two different ways of incorporating regression effects within the Kalman filter are described in Section 6.2. The standard Kalman filter recursion for the variance matrix of the filtered state vector does not rule out the possibility that this matrix becomes negative definite; this is obviously an undesirable outcome since it indicates the presence of rounding errors. The square root Kalman filter eliminates this problem at the expense of slowing down the filtering and smoothing processes; details are given in Section 6.3. The computational costs of implementing the filtering and smoothing procedures of Chapters 4 and 5 can become high for high-dimensional multivariate models, particularly in dealing with the initialisation problem. It turns out that by bringing the elements of the observation vectors in multivariate models into the filtering and smoothing computing operations one at a time, substantial gains in computational efficiency are achieved and the initialisation problem is simplified considerably. Methods based on this idea are developed in Section 6.4. The various algorithms presented in the Chapters 4 and 5 and this chapter need to be implemented efficiently on a computer. Different computer packages have been developed that implement the algorithms considered in this book. *SsfPack* is an example of such a package. Section 6.7 reviews the features of packages for state space methods and provides an illustration using *SsfPack*.

6.2 Regression estimation

6.2.1 Introduction

As for the structural time series model considered in Section 3.2, the general state space model can be extended to allow for the incorporation of explanatory variables and intervention variables into the model. To accomplish this generalisation we replace the measurement equation of the state space model (3.1) by

$$y_t = Z_t\alpha_t + X_t\beta + \varepsilon_t, \tag{6.1}$$

where $X_t = (x_{1,t}, \dots, x_{k,t})$ is a $p \times k$ matrix of explanatory variables and β is a $k \times 1$ vector of unknown regression coefficients which we assume are constant over time and which we wish to estimate. We will not discuss here time-varying regression coefficients because they can be included as part of the state vector α_t in an obvious way as in Subsection 3.6.1 and are then dealt with by the standard Kalman filter and smoother. There are two ways in which the inclusion of regression effects with fixed coefficients can be handled. First, we may include the coefficient vector β in the state vector. Alternatively, and particularly on occasions where we wish to keep the dimensionality of the state vector as low as possible, we can use the augmented Kalman filter and smoother. Both solutions will be discussed in the next two sections. Different types of residuals exist when regression variables are included in the model. We show in Subsection 6.2.4 how to calculate them within the two different solutions.

6.2.2 Inclusion of coefficient vector in state vector

The state space model in which the constant coefficient vector β in (6.1) is included in the state vector has the form

$$y_t = [Z_t \quad X_t] \begin{pmatrix} \alpha_t \\ \beta_t \end{pmatrix} + \varepsilon_t,$$

$$\begin{pmatrix} \alpha_{t+1} \\ \beta_{t+1} \end{pmatrix} = \begin{bmatrix} T_t & 0 \\ 0 & I_k \end{bmatrix} \begin{pmatrix} \alpha_t \\ \beta_t \end{pmatrix} + \begin{bmatrix} R_t \\ 0 \end{bmatrix} \eta_t,$$

for $t = 1, \dots, n$. In the initial state vector, β_1 can be taken as diffuse or fixed. In the diffuse case the model for the initial state vector is

$$\begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} a \\ 0 \end{pmatrix}, \kappa \begin{bmatrix} P_\infty & 0 \\ 0 & I_k \end{bmatrix} + \begin{bmatrix} P_* & 0 \\ 0 & 0 \end{bmatrix} \right\},$$

where $\kappa \to \infty$; see also Subsection 5.6.4 where we give the initial state vector for the regression model with ARMA errors. We attach suffixes to the β's purely for convenience in the state space formulation since $\beta_{t+1} = \beta_t = \beta$. The exact initial Kalman filter (5.19) and the Kalman filter (4.24) can be applied straightforwardly to this enlarged state space model to obtain the estimate of β. The enlargement of the state space model will not cause much extra computing because of the sparse nature of the system matrices.

6.2.3 Regression estimation by augmentation

Another method of estimating β is by augmentation of the Kalman filter. This technique is essentially the same as was used in the augmented Kalman filter of Section 5.7. We will give details of this approach on the assumption that the initial state vector does not contain diffuse elements. The likelihood function in terms of β is constructed by applying the Kalman filter to the variables y_t, $x_{1,t}, \dots, x_{k,t}$ in turn. Each of the variables $x_{1,t}, \dots, x_{k,t}$ is treated in the Kalman filter as the observation vector with the same variance elements as used for y_t. Denote the resulting one-step ahead forecast errors by v_t^*, $x_{1,t}^*, \dots, x_{k,t}^*$, respectively. Since the filtering operations are linear, the one-step ahead forecast errors for the series $y_t - X_t\beta$ are given by $v_t = v_t^* - X_t^*\beta$ where $X_t^* = (x_{1,t}^* \dots x_{k,t}^*)$. It should be noted that the $k+1$ Kalman filters are the same except that the values for v_t and a_t in (4.24) are different. We can therefore combine these filters into an augmented Kalman filter where we replace vector y_t by matrix (y_t, X_t) to obtain the 'innovations' matrix (v_t^*, X_t^*); this is analogous to the augmented Kalman filter described in Section 5.7.

The one-step ahead forecast errors $v_1, \dots, v_n$, with the corresponding variances $F_1, \dots, F_n$, are independent of each other by construction. The sum of squared standardised forecast errors become therefore simply

$$\sum_{t=1}^n v_t' F_t^{-1} v_t = \sum_{t=1}^n (v_t^* - X_t^*\beta)' F_t^{-1} (v_t^* - X_t^*\beta),$$

and can be minimised with respect to β directly. We then obtain the *generalised least squares estimate* of β with its variance matrix. They are given by

$$\hat\beta = \left(\sum_{t=1}^n X_t^{*\prime} F_t^{-1} X_t^* \right)^{-1} \sum_{t=1}^n X_t^{*\prime} F_t^{-1} v_t^*, \qquad \mathrm{Var}(\hat\beta) = \left(\sum_{t=1}^n X_t^{*\prime} F_t^{-1} X_t^* \right)^{-1}. \tag{6.2}$$

In the case where the initial state vector contains diffuse elements we can extend the augmented Kalman filter for δ as shown in Section 5.7. However, we ourselves prefer to use the exact initial Kalman filter for dealing with δ. The equations for $P_{*,t}$ and $P_{\infty,t}$ of the exact initial Kalman filter are not affected since they do not depend on the data. The update for the augmented state vector is given by the equations

$$\begin{aligned} (v_t^*, x_{1,t}^*, \dots, x_{k,t}^*) &= (y_t, x_{1,t}, \dots, x_{k,t}) - Z_t(a_{a,t}, A_{x,t}), \\ (a_{a,t+1}, A_{x,t+1}) &= T_t(a_{a,t}, A_{x,t}) + K_t^{(0)}(v_t^*, x_{1,t}^*, \dots, x_{k,t}^*), \end{aligned} \tag{6.3}$$

for $t = 1, \dots, d$ with $(a_{a,1}, A_{x,1}) = (a, 0, \dots, 0)$ and where $K_t^{(0)}$ is defined in Section 5.2. Note that F_t^{-1} in (6.2) must be replaced by zero or $F_{*,t}^{-1}$ depending on the value for $F_{\infty,t}$ in the exact initial Kalman filter. Overall, the treatment given in the previous section where we include β in the state vector, treat δ and β as diffuse and then apply the exact initial Kalman filter is conceptually simpler, though it may not be as efficient computationally for large models.

6.2.4 Least squares and recursive residuals

By considering the measurement equation (6.1) we define two different types of residuals following Harvey (1989): recursive residuals and least squares residuals. The first type are defined as

$$v_t = y_t - Z_t a_t - X_t \hat{\beta}_{t-1}, \qquad t = d+1, \dots, n,$$

where $\hat{\beta}_{t-1}$ is the maximum likelihood estimate of β given Y_{t-1}. The residuals v_t are computed easily by including β in the state vector with α_1 diffuse since the filtered state vector of the enlarged model in Subsection 6.2.2 contains $\hat{\beta}_{t-1}$. The augmentation method can of course also evaluate v_t but it needs to compute $\hat{\beta}_{t-1}$ at each time point which is not computationally efficient. Note that the residuals v_t are serially uncorrelated. The least squares residuals are given by

$$v_t^+ = y_t - Z_t a_t - X_t \hat{\beta}, \qquad t = d+1, \dots, n,$$

where $\hat{\beta}$ is the maximum likelihood estimate of β based on the entire series, so $\hat{\beta} = \hat{\beta}_n$. For the case where the method of Subsection 6.2.2 is used to compute $\hat{\beta}$, we require two Kalman filters: one for the enlarged model to compute $\hat{\beta}$ and a Kalman filter for the constructed measurement equation $y_t - X_t\hat{\beta} = Z_t\alpha_t + \varepsilon_t$ whose 'innovations' v_t are actually v_t^+ for $t = d+1, \dots, n$. The same applies to the method of Subsection 6.2.3, except that $\hat{\beta}$ is computed using equation (6.2). The least squares residuals are correlated due to the presence of $\hat{\beta}$ in these residuals, which is calculated from the whole sample.

Both sets of residuals can be used for diagnostic purposes. For these purposes the residuals v_t have the advantage of being serially uncorrelated whereas the residuals v_t^+ have the advantage of being based on the estimate $\hat{\beta}$ calculated from the whole sample. For further discussion we refer to Harvey (1989, Section 7.4.1).

6.3 Square root filter and smoother

6.3.1 Introduction

In this section we deal with the situation where, because of rounding errors and matrices being close to singularity, the possibility arises that the calculated value of P_t is negative definite, or close to this, giving rise to unacceptable rounding errors. From (4.24), the state variance matrix P_t is updated by the Kalman filter equations

$$\begin{aligned} F_t &= Z_t P_t Z_t' + H_t, \\ K_t &= T_t P_t Z_t' F_t^{-1}, \\ P_{t+1} &= T_t P_t L_t' + R_t Q_t R_t' \\ &= T_t P_t T_t' + R_t Q_t R_t' - K_t F_t K_t', \end{aligned} \tag{6.4}$$

where $L_t = T_t - K_t Z_t$. It can happen that the calculated value of P_t becomes negative definite when, for example, erratic changes occur in the system matrices over time. The problem can be avoided by using a transformed version of the Kalman filter called the *square root filter*. However, the amount of computation required is substantially larger than that required for the standard Kalman filter. The square root filter is based on orthogonal lower triangular transformations for which we can use Givens rotation techniques. The standard reference to square root filtering is Morf and Kailath (1975).

6.3.2 Square root form of variance updating

Define the partitioned matrix U_t by

$$U_t = \begin{bmatrix} Z_t\tilde{P}_t & \tilde{H}_t & 0 \\ T_t\tilde{P}_t & 0 & R_t\tilde{Q}_t \end{bmatrix}, \tag{6.5}$$

where

$$P_t = \tilde{P}_t\tilde{P}_t', \qquad H_t = \tilde{H}_t\tilde{H}_t', \qquad Q_t = \tilde{Q}_t\tilde{Q}_t',$$

in which the matrices $\tilde{P}_t$, $\tilde{H}_t$ and $\tilde{Q}_t$ are lower triangular matrices. It follows that

$$U_tU_t' = \begin{bmatrix} F_t & Z_tP_tT_t' \\ T_tP_tZ_t' & T_tP_tT_t' + R_tQ_tR_t' \end{bmatrix}. \tag{6.6}$$

The matrix U_t can be transformed to a lower triangular matrix using the orthogonal matrix G such that $GG' = I_{m+p+r}$. Note that a lower triangular matrix for a rectangular matrix such as U_t, where the number of columns exceeds the number of rows, is defined as a matrix of the form $[A \quad 0]$ where A is a square and lower triangular matrix. Postmultiplying by G we have

$$U_tG = U_t^*, \tag{6.7}$$

and $U_t^*U_t^{*\prime} = U_tU_t'$ as given by (6.6). The lower triangular rectangular matrix U_t^* has the same dimensions as U_t and can be represented as the partitioned matrix

$$U_t^* = \begin{bmatrix} U_{1,t}^* & 0 & 0 \\ U_{2,t}^* & U_{3,t}^* & 0 \end{bmatrix},$$

where $U_{1,t}^*$ and $U_{3,t}^*$ are lower triangular square matrices. It follows that

$$\begin{aligned} U_t^*U_t^{*\prime} &= \begin{bmatrix} U_{1,t}^*U_{1,t}^{*\prime} & U_{1,t}^*U_{2,t}^{*\prime} \\ U_{2,t}^*U_{1,t}^{*\prime} & U_{2,t}^*U_{2,t}^{*\prime} + U_{3,t}^*U_{3,t}^{*\prime} \end{bmatrix} \\ &= \begin{bmatrix} F_t & Z_tP_tT_t' \\ T_tP_tZ_t' & T_tP_tT_t' + R_tQ_tR_t' \end{bmatrix}, \end{aligned}$$

from which we deduce that

$$
\begin{aligned}
U^*_{1,t} &= \tilde{F}_t, \\
U^*_{2,t} &= T_t P_t Z_t' \tilde{F}_t'^{-1} = K_t \tilde{F}_t,
\end{aligned}
$$

where $F_t = \tilde{F}_t \tilde{F}_t'$ and $\tilde{F}_t$ is lower triangular. It is remarkable to find that $U^*_{3,t} = \tilde{P}_{t+1}$ since

$$
\begin{aligned}
U^*_{3,t} U^{*\prime}_{3,t} &= T_t P_t T_t' + R_t Q_t R_t' - U^*_{2,t} U^{*\prime}_{2,t} \\
&= T_t P_t T_t' + R_t Q_t R_t' - K_t F_t K_t' \\
&= P_{t+1},
\end{aligned}
$$

which follows from (6.4). Thus by transforming U_t in (6.5) to a lower triangular matrix we obtain $\tilde{P}_{t+1}$; this operation can thus be regarded as a square root recursion for P_t. The update for the state vector a_t can be easily incorporated using

$$
\begin{aligned}
a_{t+1} &= T_t a_t + K_t v_t, \\
&= T_t a_t + T_t P_t Z_t' \tilde{F}_t'^{-1} \tilde{F}_t^{-1} v_t \\
&= T_t a_t + U^*_{2,t} U^{*-1}_{1,t} v_t,
\end{aligned}
$$

where $v_t = y_t - Z_t a_t$. Note that the inverse of $U^*_{1,t}$ is easy to calculate since it is a lower triangular matrix.

6.3.3 Givens rotations

Matrix G can be any orthogonal matrix which transforms U_t to a lower triangular matrix. Many different techniques can be used to achieve this objective; for example, Golub and Van Loan (1996) give a detailed treatment of the Householder and Givens matrices for the purpose. We will give here a short description of the latter. The orthogonal 2×2 matrix

$$
G_2 = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}, \qquad (6.8)
$$

with $c^2 + s^2 = 1$ is the key to Givens transformations. It is used to transform the vector

$$
x = (x_1 \quad x_2),
$$

into a vector in which the second element is zero, that is

$$
y = x G_2 = (y_1 \quad 0),
$$

by taking

$$
c = \frac{x_1}{\sqrt{x_1^2 + x_2^2}}, \qquad s = -\frac{x_2}{\sqrt{x_1^2 + x_2^2}}, \qquad (6.9)
$$

for which $c^2 + s^2 = 1$ and $sx_1 + cx_2 = 0$. Note that $y_1 = cx_1 - sx_2$ and $yG_2' = xG_2G_2' = x$.

The general Givens matrix G is defined as the identity matrix I_q but with four elements $I_{ii}, I_{jj}, I_{ij}, I_{ji}$ replaced by

$$\begin{aligned} G_{ii} &= G_{jj} = c, \\ G_{ij} &= s, \\ G_{ji} &= -s, \end{aligned}$$

for $1 \leq i < j \leq q$ and with c and s given by (6.9) but now enforcing element (i, j) of matrix XG to be zero for all $1 \leq i < j \leq q$ and for any matrix X. It follows that $GG' = I$ so when Givens rotations are applied repeatedly to create zero blocks in a matrix, the overall transformation matrix is also orthogonal. These properties of the Givens rotations, their computational efficiency and their numerical stability make them a popular tool to transform nonzero matrices into sparse matrices such as a lower triangular matrix. More details and efficient algorithms for Givens rotations are given by Golub and Van Loan (1996).

6.3.4 Square root smoothing

The backward recursion (4.42) for N_{t-1} of the basic smoothing equations can also be given in a square root form. These equations use the output of the square root Kalman filter as given by:

$$\begin{aligned} U_{1,t}^* &= \tilde{F}_t, \\ U_{2,t}^* &= K_t\tilde{F}_t, \\ U_{3,t}^* &= \tilde{P}_{t+1}. \end{aligned}$$

The recursion for N_{t-1} is given by

$$N_{t-1} = Z_t'F_t^{-1}Z_t + L_t'N_tL_t,$$

where

$$\begin{aligned} F_t^{-1} &= (U_{1,t}^*U_{1,t}^{*\prime})^{-1}, \\ L_t &= T_t - U_{2,t}^*U_{1,t}^{*-1}Z_t. \end{aligned}$$

We introduce the lower triangular square matrix $\tilde{N}_t$ such that

$$N_t = \tilde{N}_t\tilde{N}_t',$$

and the $m \times (m + p)$ matrix

$$\tilde{N}_{t-1}^* = \left[\begin{array}{cc} Z_t'U_{1,t}^{*-1\prime} & L_t'\tilde{N}_t \end{array}\right],$$

from which it follows that $N_{t-1} = \tilde{N}_{t-1}^*\tilde{N}_{t-1}^{*\prime}$.

The matrix N^*_{t-1} can be transformed to a lower triangular matrix using some orthogonal matrix G such that $GG' = I_{m+p}$; compare Subsection 6.3.2. We have

$$\tilde{N}^*_{t-1} G = [\, \tilde{N}_{t-1} \quad 0\,],$$

such that $N_{t-1} = \tilde{N}^*_{t-1}\tilde{N}^{*\prime}_{t-1} = \tilde{N}_{t-1}\tilde{N}'_{t-1}$. Thus by transforming the matrix N^*_{t-1}, depending on matrices indexed by time t, to a lower triangular matrix we obtain the square root matrix of N_{t-1}. Consequently we have developed a backward recursion for N_{t-1} in square root form. The backwards recursion for r_{t-1} is not affected apart from the way in which F_t^{-1} and L_t are computed.

6.3.5 Square root filtering and initialisation

The square root formulation could be developed for the exact initial version of the Kalman filter of Section 5.2. However, the motivation for developing square root versions for filtering and smoothing is to avoid computational numerical instabilities due to rounding errors which are built up during the recursive computations. Since initialisation usually only requires a limited number of d updates, the numerical problems are not substantial during this process. Thus, although use of the square root filter may be important for $t = d, \ldots, n$, it will normally be adequate to employ the standard exact initial Kalman filter as described in Section 5.2 for $t = 1, \ldots, d$.

The square root formulation of the augmented Kalman filter of Section 5.7 is more or less the same as the usual Kalman filter because updating equations for F_t, K_t and P_t are unaltered. Some adjustments are required for the updating of the augmented quantities but these can be derived straightforwardly; some details are given by de Jong (1991). Snyder and Saligari (1996) have proposed a Kalman filter based on Givens rotations, such as the ones developed in Subsection 6.3.3, with the fortunate property that diffuse priors $\kappa \to \infty$ can be dealt with explicitly within the Givens operations. Their application of this solution, however, was limited to filtering only and it does not seem to provide an adequate solution for initial diffuse smoothing.

6.3.6 Illustration: local linear trend model

For the local linear trend model (3.2) we take

$$U_t = \begin{bmatrix} \tilde{P}_{11,t} & 0 & \sigma_\varepsilon & 0 & 0 \\ \tilde{P}_{11,t} + \tilde{P}_{21,t} & \tilde{P}_{22,t} & 0 & \sigma_\xi & 0 \\ \tilde{P}_{21,t} & \tilde{P}_{22,t} & 0 & 0 & \sigma_\zeta \end{bmatrix},$$

which is transformed to the lower triangular matrix

$$U^*_t = \begin{bmatrix} \tilde{F}_t & 0 & 0 & 0 & 0 \\ K_{11,t}\tilde{F}_t & \tilde{P}_{11,t+1} & 0 & 0 & 0 \\ K_{21,t}\tilde{F}_t & \tilde{P}_{21,t+1} & \tilde{P}_{22,t+1} & 0 & 0 \end{bmatrix}.$$

The zero elements of U_t^* are created row-wise by a sequence of Givens rotations applied to the matrix U_t. Some zero elements in U_t^* are already zero in U_t and they mostly remain zero within the overall Givens transformation so the number of computations can be limited somewhat.

6.4 Univariate treatment of multivariate series

6.4.1 Introduction

In Chapters 4 and 5 and in this chapter we have treated the filtering and smoothing of multivariate series in the traditional way by taking the entire observational vectors y_t as the items for analysis. In this section we present an alternative approach in which the elements of y_t are brought into the analysis one at a time, thus in effect converting the multivariate series into a univariate time series. This device not only offers significant computational gains for the filtering and smoothing of the bulk of the series but it also provides substantial simplification of the initialisation process when the initial state vector α_1 is partially or wholly diffuse.

This univariate approach to vector observations was suggested for filtering by Anderson and Moore (1979) and for filtering and smoothing longitudinal models by Fahrmeir and Tutz (1994). The treatment given by these authors was, however, incomplete and in particular did not deal with the initialisation problem, where the most substantial gains are made. The following discussion of the univariate approach is based on Koopman and Durbin (2000) who gave a complete treatment including a discussion of the initialisation problem.

6.4.2 Details of univariate treatment

Our analysis will be based on the standard model

$$y_t = Z_t\alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = T_t\alpha_t + R_t\eta_t,$$

with $\varepsilon_t \sim \mathrm{N}(0, H_t)$ and $\eta_t \sim \mathrm{N}(0, Q_t)$ for $t = 1, \ldots, n$. To begin with, let us assume that $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ and H_t is diagonal; this latter restriction will be removed later. On the other hand, we introduce two slight generalisations of the basic model: first, we permit the dimensionality of y_t to vary over time by taking the dimension of vector y_t to be $p_t \times 1$ for $t = 1, \ldots, n$; second, we do not require the prediction error variance matrix F_t to be nonsingular.

Write the observation and disturbance vectors as

$$y_t = \begin{pmatrix} y_{t,1} \\ \vdots \\ y_{t,p_t} \end{pmatrix}, \qquad \varepsilon_t = \begin{pmatrix} \varepsilon_{t,1} \\ \vdots \\ \varepsilon_{t,p_t} \end{pmatrix},$$

and the observation equation matrices as

$$Z_t = \begin{pmatrix} Z_{t,1} \\ \vdots \\ Z_{t,p_t} \end{pmatrix}, \qquad H_t = \begin{pmatrix} \sigma_{t,1}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_{t,p_t}^2 \end{pmatrix},$$

where $y_{t,i}$, $\varepsilon_{t,i}$ and $\sigma_{t,i}^2$ are scalars and $Z_{t,i}$ is a $(1 \times m)$ row vector, for $i = 1, \dots, p_t$. The observation equation for the univariate representation of the model is

$$y_{t,i} = Z_{t,i}\alpha_{t,i} + \varepsilon_{t,i}, \qquad i = 1, \dots, p_t, \qquad t = 1, \dots, n, \tag{6.10}$$

where $\alpha_{t,i} = \alpha_t$. The state equation corresponding to (6.10) is

$$\begin{aligned} \alpha_{t,i+1} &= \alpha_{t,i}, & i &= 1, \dots, p_t - 1, \\ \alpha_{t+1,1} &= T_t\alpha_{t,p_t} + R_t\eta_t, & t &= 1, \dots, n, \end{aligned} \tag{6.11}$$

where the initial state vector $\alpha_{1,1} = \alpha_1 \sim N(a_1, P_1)$.

Define

$$a_{t,1} = \mathrm{E}(\alpha_{t,1}|Y_{t-1}), \qquad P_{t,1} = \mathrm{Var}(\alpha_{t,1}|Y_{t-1}),$$

and

$$\begin{aligned} a_{t,i} &= \mathrm{E}(\alpha_{t,i}|Y_{t-1}, y_{t,1}, \dots, y_{t,i-1}), \\ P_{t,i} &= \mathrm{Var}(\alpha_{t,i}|Y_{t-1}, y_{t,1}, \dots, y_{t,i-1}), \end{aligned}$$

for $i = 2, \dots, p_t$. By treating the vector series $y_1, \dots, y_n$ as the scalar series

$$y_{1,1}, \dots, y_{1,p_1}, y_{2,1}, \dots, y_{n,p_n},$$

the filtering equations (4.24) can be written as

$$a_{t,i+1} = a_{t,i} + K_{t,i}v_{t,i}, \qquad P_{t,i+1} = P_{t,i} - K_{t,i}F_{t,i}K_{t,i}', \tag{6.12}$$

where

$$v_{t,i} = y_{t,i} - Z_{t,i}a_{t,i}, \qquad F_{t,i} = Z_{t,i}P_{t,i}Z_{t,i}' + \sigma_{t,i}^2, \qquad K_{t,i} = P_{t,i}Z_{t,i}'F_{t,i}^{-1}, \tag{6.13}$$

for $i = 1, \dots, p_t$ and $t = 1, \dots, n$. This formulation has $v_{t,i}$ and $F_{t,i}$ as scalars and $K_{t,i}$ as a column vector. The transition from time t to time $t + 1$ is achieved by the relations

$$a_{t+1,1} = T_t a_{t,p_t+1}, \qquad P_{t+1,1} = T_t P_{t,p_t+1} T_t' + R_t Q_t R_t'. \tag{6.14}$$

The values $a_{t+1,1}$ and $P_{t+1,1}$ are the same as the values a_{t+1} and P_{t+1} computed by the standard Kalman filter.

It is important to note that the elements of the innovation vector v_t are not the same as $v_{t,i}$, for $i = 1, \dots, p_t$; only the first element of v_t is equal to $v_{t,1}$. The same applies to the diagonal elements of the variance matrix F_t and the variances $F_{t,i}$, for $i = 1, \dots, p_t$; only the first diagonal element of F_t is equal to $F_{t,1}$. It should be emphasised that there are models for which $F_{t,i}$ can be zero, for example the case where y_t is a multinomial observation with all cell counts included in y_t. This indicates that $y_{t,i}$ is linearly dependent on previous observations for some i. In this case,

$$\begin{aligned} a_{t,i+1} &= \mathrm{E}(\alpha_{t,i+1}|Y_{t-1}, y_{t,1}, \dots, y_{t,i}) \\ &= \mathrm{E}(\alpha_{t,i+1}|Y_{t-1}, y_{t,1}, \dots, y_{t,i-1}) = a_{t,i}, \end{aligned}$$

and similarly $P_{t,i+1} = P_{t,i}$. The contingency is therefore easily dealt with.

The basic smoothing recursions (4.69) for the standard state space model can be reformulated for the univariate series

$$y_{1,1}, \dots, y_{1,p_t}, y_{2,1}, \dots, y_{n,p_n},$$

as

$$\begin{aligned} r_{t,i-1} &= Z'_{t,i}F^{-1}_{t,i}v_{t,i} + L'_{t,i}r_{t,i}, & N_{t,i-1} &= Z'_{t,i}F^{-1}_{t,i}Z_{t,i} + L'_{t,i}N_{t,i}L_{t,i}, \\ r_{t-1,p_{t-1}} &= T'_{t-1}r_{t,0}, & N_{t-1,p_{t-1}} &= T'_{t-1}N_{t,0}T_{t-1}, \end{aligned} \tag{6.15}$$

where $L_{t,i} = I_m - K_{t,i}Z_{t,i}$, for $i = p_t, \dots, 1$ and $t = n, \dots, 1$. The initialisations are $r_{n,p_n} = 0$ and $N_{n,p_n} = 0$. The equations for $r_{t-1,p_{t-1}}$ and $N_{t-1,p_{t-1}}$ do not apply for $t = 1$. The values for $r_{t,0}$ and $N_{t,0}$ are the same as the values for the smoothing quantities r_{t-1} and N_{t-1} of the standard smoothing equations, respectively.

The smoothed state vector $\hat{\alpha}_t = \mathrm{E}(\alpha_t|Y_n)$ and the variance error matrix $V_t = \mathrm{Var}(\alpha_t|Y_n)$, together with other related smoothing results for the transition equation, are computed by using the standard equations (4.39) and (4.43) with

$$a_t = a_{t,1}, \qquad P_t = P_{t,1}, \qquad r_{t-1} = r_{t,0}, \qquad N_{t-1} = N_{t,0}.$$

Finally, the smoothed estimators for the observation disturbances $\varepsilon_{t,i}$ of (6.10) follow directly from our approach and are given by

$$\begin{aligned} \widehat{\varepsilon}_{t,i} &= \sigma^2_{t,i}F^{-1}_{t,i}(v_{t,i} - K'_{t,i}r_{t,i}), \\ \mathrm{Var}(\,\widehat{\varepsilon}_{t,i}) &= \sigma^4_{t,i}F^{-2}_{t,i}(F_{t,i} + K'_{t,i}N_{t,i}K_{t,i}). \end{aligned}$$

Since the simulation smoothers of Subsection 4.9.1 and 4.9.2 rely fully on the application of the Kalman filter and smoother, the univariate treatment of multivariate observations does not have further consequences for simulation smoothing.

6.4.3 Correlation between observation equations

For the case where H_t is not diagonal, the univariate representation of the state space model (6.10) does not apply due to the correlations between the $\varepsilon_{t,i}$'s. In this situation we can pursue two different approaches. First, we can put the disturbance vector ε_t into the state vector. For the observation equation of (3.1) define

$$\bar{\alpha}_t = \begin{pmatrix} \alpha_t \\ \varepsilon_t \end{pmatrix}, \qquad \bar{Z}_t = \begin{pmatrix} Z_t & I_{P_t} \end{pmatrix},$$

and for the state equation define

$$\bar{\eta}_t = \begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix}, \qquad \bar{T}_t = \begin{pmatrix} T_t & 0 \\ 0 & 0 \end{pmatrix},$$

$$\bar{R}_t = \begin{pmatrix} R_t & 0 \\ 0 & I_{P_t} \end{pmatrix}, \qquad \bar{Q}_t = \begin{pmatrix} Q_t & 0 \\ 0 & H_t \end{pmatrix},$$

leading to

$$y_t = \bar{Z}_t \bar{\alpha}_t, \qquad \bar{\alpha}_{t+1} = \bar{T}_t \bar{\alpha}_t + \bar{R}_t \bar{\eta}_t, \qquad \bar{\eta}_t \sim \mathrm{N}(0, \bar{Q}_t),$$

for $t = 1, \ldots, n$. We then proceed with the same technique as for the case where H_t is diagonal by treating each element of the observation vector individually. The second approach is to transform the observations. In the case where H_t is not diagonal, we diagonalise it by the Cholesky decomposition

$$H_t = C_t H_t^* C_t',$$

where H_t^* is diagonal and C_t is lower triangular with ones on the diagonal. By transforming the observations, we obtain the observation equation

$$y_t^* = Z_t^* \alpha_t + \varepsilon_t^*, \qquad \varepsilon_t^* \sim \mathrm{N}(0, H_t^*),$$

where $y_t^* = C_t^{-1} y_t$, $Z_t^* = C_t^{-1} Z_t$ and $\varepsilon_t^* = C_t^{-1} \varepsilon_t$. Since C_t is a lower triangular matrix, it is easy to compute its inverse. The state vector is not affected by the transformation. Since the elements of ε_t^* are independent we can treat the series y_t^* as a univariate series in the above way.

These two approaches for correlated observation disturbances are complementary. The first method has the drawback that the state vector can become large. The second method is illustrated in Subsection 6.4.5 where we show that simultaneously transforming the state vector can also be convenient.

6.4.4 Computational efficiency

The main motivation for this 'univariate' approach to filtering and smoothing for multivariate state space models is computational efficiency. This approach avoids the inversion of matrix F_t and two matrix multiplications. Also, the implementation of the recursions is more straightforward. Table 6.1 shows that the percentage savings in the number multiplications for filtering using the univariate approach compared to the standard approach are considerable. The calculations concerning the transition are not taken into account in the calculations for this table because matrix T_t is usually sparse with most elements equal to zero or unity.

Table 6.2 presents the considerable percentage savings in the number of multiplications for state smoothing compared to the standard multivariate approach. Again, the computations involving the transition matrix T_t are not taken into account in compiling these figures.

6.4.5 Illustration: vector splines

We now consider the application of the univariate approach to vector splines. The generalisation of smoothing splines of Hastie and Tibshirani (1990) to the multivariate case is considered by Fessler (1991) and Yee and Wild (1996). The vector spline model is given by

$$y_i = \theta(t_i) + \varepsilon_i, \qquad \mathrm{E}(\varepsilon_i) = 0, \qquad \mathrm{Var}(\varepsilon_i) = \Sigma_i, \qquad i = 1, \dots, n,$$

where y_i is a $p \times 1$ vector response at scalar t_i, $\theta(\cdot)$ is an arbitrary smooth vector function and errors ε_i are mutually uncorrelated. The variance matrix Σ_i is assumed to be known and is usually constant for varying i. This is a

Table 6.1 Percentage savings for filtering using univariate approach.

State dim.	Obs. dim.	$p = 1$	$p = 2$	$p = 3$	$p = 5$	$p = 10$	$p = 20$
$m = 1$		0	39	61	81	94	98
$m = 2$		0	27	47	69	89	97
$m = 3$		0	21	38	60	83	95
$m = 5$		0	15	27	47	73	90
$m = 10$		0	8	16	30	54	78
$m = 20$		0	5	9	17	35	58

Table 6.2 Percentage savings for smoothing using univariate approach.

State dim.	Obs. dim.	$p = 1$	$p = 2$	$p = 3$	$p = 5$	$p = 10$	$p = 20$
$m = 1$		0	27	43	60	77	87
$m = 2$		0	22	36	53	72	84
$m = 3$		0	19	32	48	68	81
$m = 5$		0	14	25	40	60	76
$m = 10$		0	9	16	28	47	65
$m = 20$		0	5	10	18	33	51

generalisation of the univariate problem considered in Subsection 3.9.2. The standard method of estimating the smooth vector function is by minimising the generalised least squares criterion

$$\sum_{i=1}^{n} \{y_i - \theta(t_i)\}' \Sigma_i^{-1} \{y_i - \theta(t_i)\} + \sum_{j=1}^{p} \lambda_j \int \theta_j''(t)^2 \, dt,$$

where the non-negative smoothing parameter λ_j determines the smoothness of the jth smooth function $\theta_j(\cdot)$ of vector $\theta(\cdot)$ for $j = 1, \ldots, p$. Note that $t_{i+1} > t_i$ for $i = 1, \ldots, n-1$ and $\theta_j''(t)$ denotes the second derivative of $\theta_j(t)$ with respect to t. In the same way as for the univariate case in (3.41), we use the discrete model

$$\begin{aligned}
y_i &= \mu_i + \varepsilon_i, \\
\mu_{i+1} &= \mu_i + \delta_i \nu_i + \eta_i, \qquad & \mathrm{Var}(\eta_i) &= \frac{\delta_i^3}{3}\Lambda, \\
\nu_{i+1} &= \nu_i + \zeta_i, \qquad & \mathrm{Var}(\zeta_i) &= \delta_i \Lambda, \qquad \mathrm{Cov}(\eta_i, \zeta_i) = \frac{\delta_i^2}{2}\Lambda,
\end{aligned}$$

with vector $\mu_i = \theta(t_i)$, scalar $\delta_i = t_{i+1} - t_i$ and diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. This model is the multivariate extension of the continuous-time representation of the integrated random walk model that is introduced in Subsection 3.2.1; see Harvey (1989, p. 487). In the case of $\Sigma_i = \Sigma$ and with the Cholesky decomposition $\Sigma = CDC'$ where matrix C is lower triangular and matrix D is diagonal, we obtain the transformed model

$$\begin{aligned}
y_i^* &= \mu_i^* + \varepsilon_i^*, \\
\mu_{i+1}^* &= \mu_i^* + \delta_i \nu_i^* + \eta_i^*, \qquad & \mathrm{Var}(\eta_i) &= \frac{\delta_i^3}{3}Q, \\
\nu_{i+1}^* &= \nu_i^* + \zeta_i^*, \qquad & \mathrm{Var}(\zeta_i) &= \delta_i Q, \qquad \mathrm{Cov}(\eta_i, \zeta_i) = \frac{\delta_i^2}{2}Q,
\end{aligned}$$

with $y_i^* = C^{-1} y_i$ and $\mathrm{Var}(\varepsilon_i^*) = D$ where we have used (3.42). Furthermore, we have $\mu_i^* = C^{-1}\mu_i$, $\nu_i^* = C^{-1}\nu_i$ and $Q = C^{-1}\Lambda C'^{-1}$. The Kalman filter smoother algorithm provides the fitted smoothing spline. The untransformed model and the transformed model can both be handled by the univariate strategy of filtering and smoothing discussed in this section. The advantage of the transformed model is that ε_i^* can be excluded from the state vector which is not possible for the untransformed model because $\mathrm{Var}(\varepsilon_i) = \Sigma_i$ is not necessarily diagonal; see the discussion in Subsection 6.4.3.

The percentage computational saving of the univariate approach for spline smoothing depends on the size p. The state vector dimension for the transformed model is $m = 2p$ so the percentage saving in computing for filtering is 30 if $p = 5$ and it is 35 if $p = 10$; see Table 6.1. The percentages for smoothing are 28 and 33, respectively; see Table 6.2.

6.5 Collapsing large observation vectors

6.5.1 Introduction

In cases where the dimensionality p of the observation vector y_t is large in relation to the dimensionality m of the state vector in the state space model (3.1), the Kalman filter may become computationally onerous. In particular, the computation of the inverse of the $p \times p$ innovation variance matrix, F_t^{-1} in the Kalman filter (4.24) for each t can become a computational burden. The univariate treatment of the Kalman filter, as discussed in the previous section, can alleviate a significant part of this computational problem since we update the state vector estimate for each individual element in y_t and hence the innovation variance is a scalar. Nevertheless, the computations required to update the state estimate for each element of the observation vector can become burdensome when p is as large as, say, 500.

An alternative way to deal with the problem is to adopt the well-known matrix identity for the inverse of F_t,

$$F_t^{-1} = \left(Z_t P_t Z_t' + H_t\right)^{-1} = H_t^{-1} - H_t^{-1} Z_t \left(P_t^{-1} + Z_t' H_t^{-1} Z_t\right)^{-1} Z_t' H_t^{-1}, \quad (6.16)$$

which is valid when H_t and P_t are both nonsingular matrices; see, for example, Problem 2.9 of Chapter 1 in Rao (1973). Although H_t has the same large dimensionality as F_t, we can expect that matrix H_t has a convenient structure; for example, when H_t is a diagonal matrix. In cases where p is much larger than m, computing the inverse of P_t in (6.16) should be a much lighter operation than inverting F_t directly; for example, when, say, $p = 100$ and $m = 5$. Such dimensionalities are typical in the case of dynamic factors models which are discussed in Section 3.7. The drawback of using (6.16) for the computation of F_t^{-1} is the set of conditions required for H_t and P_t; they must be both nonsingular and H_t must have a convenient structure. Many models of interest do not have a nonsingular matrix H_t and/or it cannot be guaranteed that matrix P_t is nonsingular for all t. Furthermore, the univariate treatment cannot be employed once we adopt the identity (6.16) for inverting F_t.

However, a new approach has recently been developed by Jungbacker and Koopman (2008, Section 2) that collapses the original $p \times 1$ observation vector into a new observation vector of order $m \times 1$. The new observation equation can then be combined with the original state equation to provide a valid state space model. The Kalman filter and smoother for this model produces estimates which are identical to those produced by the original model. It is evident when p is very large relative to m, the use of this approach can lead to substantial computational savings; for example, when the analysis is applied to high-dimensional dynamic factor models where dimensionality of y_t can be as large as 500. Details of the approach are given in the next subsection together with illustrations.

6.5.2 Collapse by transformation

The idea behind collapsing by transformation is to transform the observation vector into a pair of vectors y_t^* and y_t^+ such that y_t^* is $m \times 1$ and y_t^+ is $(p-m) \times 1$, y_t^* depends on α_t and y_t^+ does not depend on α_t and y_t^* and y_t^+ are independent given α_t.We then replace the observation equation for y_t in the original state space model by a new observation equation for y_t^*; this gives a smaller model for analysis. We say that we have *collapsed* y_t into y_t^*.

Let A_t^* be the *projection matrix* $\left(Z_t' H_t^{-1} Z_t\right)^{-1} Z_t' H_t^{-1}$ and let $y_t^* = A_t^* y_t$. Then y_t^* is the generalised least squares estimate of α_t in the conditional distribution of α_t given y_t. Let $A_t^+ = B_t(I_p - Z_t A_t^*)$ where $(p-m) \times p$ matrix B_t is chosen such that A_t^+ has full rank $p - m$. Methods are available for calculating values of B_t with this property, but as will be seen below, we do not need to have an explicit value for B_t. It follows that $A_t^* Z_t = I_p$ and $A_t^+ Z_t = 0$. The original observation equation is given by

$$y_t = Z_t \alpha_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, H_t), \tag{6.17}$$

and the observation equation for the transformed variables is given by

$$\begin{pmatrix} y_t^* \\ y_t^+ \end{pmatrix} = \begin{bmatrix} A_t^* \\ A_t^+ \end{bmatrix} y_t = \begin{pmatrix} \alpha_t \\ 0 \end{pmatrix} + \begin{pmatrix} \varepsilon_t^* \\ \varepsilon_t^+ \end{pmatrix},$$

where $y_t^+ = A_t^+ y_t$, $\varepsilon_t^* = A_t^* \varepsilon_t$ and $\varepsilon_t^+ = A_t^+ \varepsilon_t$, for $t = 1, \ldots, n$. Since

$$\mathrm{Cov}(\varepsilon_t^*, \varepsilon_t^+) = \mathrm{E}(\varepsilon_t^* \varepsilon_t^{+\prime}) = A_t^* H_t (I_p - A_t^{*\prime} Z_t') B_t' = (A_t^* - A_t^*)\, H_t B_t' = 0,$$

the model equations for the transformed observations are given by

$$\begin{aligned} y_t^* &= \alpha_t + \varepsilon_t^*, & \varepsilon_t^* &\sim \mathrm{N}(0, H_t^*), \\ y_t^+ &= \varepsilon_t^+, & \varepsilon_t^+ &\sim \mathrm{N}(0, H_t^+), \end{aligned}$$

where ε_t^* and ε_t^+ are independent with variance matrices $H_t^* = A_t^* H_t A_t^{*\prime}$ and $H_t^+ = A_t^+ H_t A_t^{+\prime}$, respectively. All information about α_t is contained in the first equation and therefore we can discard the second equation and we can replace the original state space model by the collapsed model

$$\begin{aligned} y_t^* &= \alpha_t + \varepsilon_t^*, & \varepsilon_t^* &\sim \mathrm{N}(0, H_t^*), \\ \alpha_{t+1} &= T_t \alpha_t + R_t \eta_t, & \eta_t &\sim \mathrm{N}(0, Q_t), \end{aligned} \tag{6.18}$$

for $t = 1, \ldots, n$ where y_t^* is $m \times 1$. Analysis of this model by the methods of Chapter 4 gives exactly the same results for Kalman filtering and smoothing as analysis of the original model.

When p is substantially larger than m, the use of model (6.18) can therefore result in a significant saving in computation. When the original model is time-invariant in which the system matrices are constant over time, collapsing is particularly advantageous since we only need to compute the matrices A_t^* and H_t^* once.

6.5.3 A generalisation of collapsing by transformation

We now present a generalisation of the method of collapsing described in Subsection 6.5.2. Let

$$\bar{A}_t^* = C_t Z_t' H_t^{-1}, \tag{6.19}$$

where C_t is an arbitrary nonsingular and nonrandom matrix and let

$$\bar{A}_t^+ = B_t \left[I_p - Z_t \left(Z_t' H_t^{-1} Z_t \right)^{-1} Z_t' H_t^{-1} \right],$$

where B_t is the same as in Subsection 6.5.2. It follows that $\bar{A}_t^*$ is a generalisation of A_t^* of Subsection 6.5.2 where $\bar{A}_t^* = A_t^*$ when $C_t = \left(Z_t' H_t^{-1} Z_t \right)^{-1}$. We have $\bar{A}_t^+ Z_t = 0$ and consider the transformation

$$\begin{pmatrix} \bar{y}_t^* \\ \bar{y}_t^+ \end{pmatrix} = \begin{bmatrix} \bar{A}_t^* \\ \bar{A}_t^+ \end{bmatrix} y_t = \begin{pmatrix} Z_t^* \alpha_t \\ 0 \end{pmatrix} + \begin{pmatrix} \bar{\varepsilon}_t^* \\ \bar{\varepsilon}_t^+ \end{pmatrix},$$

where $Z_t^* = \bar{A}_t^* Z_t = C_t Z_t' H_t^{-1} Z_t$, $\bar{\varepsilon}_t^* = \bar{A}_t^* \varepsilon_t$ and $\bar{\varepsilon}_t^+ = \bar{A}_t^+ \varepsilon_t$, for $t = 1, \ldots, n$. Since,

$$\begin{aligned} \mathrm{E}(\bar{\varepsilon}_t^*, \bar{\varepsilon}_t^{+\prime}) &= \bar{A}_t^* H_t \left[I_p - H_t^{-1} Z_t \left(Z_t' H_t^{-1} Z_t \right)^{-1} Z_t' \right] B_t' \\ &= \left[C_t Z_t' - C_t Z_t' H_t^{-1} Z_t \left(Z_t' H_t^{-1} Z_t \right)^{-1} Z_t' \right] \; = \; 0, \end{aligned}$$

$\bar{\varepsilon}_t^*$ and $\bar{\varepsilon}_t^+$ are independent. We therefore can replace the original $p \times 1$ observation equation by the $m \times 1$ collapsed observation equation

$$\bar{y}_t^* = Z_t^* \alpha_t + \bar{\varepsilon}_t^*, \qquad \bar{\varepsilon}_t^* \sim \mathrm{N}(0, \bar{H}_t^*),$$

for $t = 1, \ldots, n$ where $\bar{H}_t^* = C_t Z_t' H_t^{-1} Z_t C_t'$. In particular, if we take $C_t = I_p$ then $\bar{A}_t = Z_t' H_t^{-1}$ and $\bar{H}_t^* = Z_t' H_t^{-1} Z_t$. This shows that the transformation is flexible and that C_t can be chosen for convenience (within the set of nonsingular matrices) depending on the details of the particular state space model under consideration.

A particularly convenient choice for C_t is to take $C_t' C_t = \left(Z_t' H_t^{-1} Z_t \right)^{-1}$ such that

$$Z_t^* = \bar{A}_t^* Z_t = C_t Z_t' H_t^{-1} Z_t = C_t (C_t' C_t)^{-1} = C_t C_t^{-1} C_t'^{\,-1} = C_t'^{\,-1},$$

and

$$\bar{H}_t^* = C_t Z_t' H_t^{-1} Z_t C_t' = C_t (C_t' C_t)^{-1} C_t' = I_p.$$

Kalman filtering and smoothing applied to the collapsed model for $\bar{y}_t^*$ with $\bar{H}_t^* = I_p$ can directly be based on the univariate treatment of Section 6.4.

The results into this section were first obtained by Jungbacker and Koopman (2008) in a context more general than considered here. For example, they show how the collapse of y_t can take place to the lowest possible dimension for $\bar{y}_t^*$.

Table 6.3 Percentage savings for Kalman filtering.

State dim.	Obs. dim.	$p = 10$	$p = 50$	$p = 100$	$p = 250$	$p = 500$
$m = 1$		50	82	85	89	92
$m = 5$		23	79	87	93	94
$m = 10$		–	68	82	92	95
$m = 25$		–	33	60	81	90
$m = 50$		–	–	33	67	81

6.5.4 Computational efficiency

To obtain some insight into the size of the computational gains due to collapsing, we apply the Kalman filter for different values of p and m, with and without collapsing the observation vector y_t to y_t^*. The percentage savings achieved by the Kalman filter for the collapsed observation vector compared to the Kalman filter for the original observation vector are reported in Table 6.3 for different dimensions p and m. The computations associated with the transformation are taken into account as well as the computations associated with the transition equation. The percentage savings of the collapsing method are considerable by any means but particularly when $p >> m$.

6.6 Filtering and smoothing under linear restrictions

We now consider how to carry out filtering and smoothing subject to a set of time-varying linear restrictions on the state vector of the form

$$R_t^* \alpha_t = r_t^*, \qquad t = 1, \dots, n, \tag{6.20}$$

where the matrix R_t^* and the vector r_t^* are known and where the number of rows in R_t^* can vary with t. Although linear restrictions on the state vector can often be easily dealt with by respecifying the elements of the state vector, an alternative is to proceed as follows. To impose the restrictions (6.20) we augment the obervation equation as

$$\begin{pmatrix} y_t \\ r_t^* \end{pmatrix} = \begin{bmatrix} Z_t \\ R_t^* \end{bmatrix} \alpha_t + \begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix}, \qquad t = 1, \dots, n. \tag{6.21}$$

For this augmented model, filtering and smoothing will produce estimates a_t and $\hat{\alpha}_t$ which are subject to the restrictions $R_t^* a_t = r_t^*$ and $R_t^* \hat{\alpha}_t = r_t^*$; for a discussion of this procedure, see Doran (1992). Equation (6.21) represents a multivariate model whether y_t is univariate or not. This can, however, be treated by a univariate approach to filtering and smoothing based on the devices discussed in Section 6.4.

6.7 Computer packages for state space methods

6.7.1 Introduction

The use of state space methods in empirical work is enhanced with the implementations of state space methods in statistical and econometric computer packages. The program *STAMP* (Structural Time Series Analyser, Modeller and Predictor) firstly appeared in 1982 and was developed by Simon Peters and Andrew Harvey at the London School of Economics. It has been widely acknowledged as one of the first statistical software packages primarily based on state space methods. The *STAMP* program includes options for maximum likelihood estimation, diagnostic checking, signal extraction and forecasting procedures applied to the structural time series models which we discuss in Section 3.2 and 3.3. *STAMP* has made these functions available in a user-friendly menu system. The more recent versions of the program include extended facilities for analysing univariate and multivariate models, automatic outlier and break detection, and forecasting. The current version is developed by Koopman, Harvey, Doornik and Shephard (2010) and the latest information can be obtained from the Internet at `http://stamp-software.com/`.

Software implementations of state space methods have increased with the progress in computing capabilities and the modernisation of computer software generally. Many well-known statistical and econometric software packages currently have options for the use of state space methods. A complete overview of available computer packages with options for using state space methods is provided in a special issue of the *Journal of Statistical Software* and is edited by Commandeur, Koopman and Ooms (2011). It shows that many software tools for state space models are currently available for the analysis of time series. An example of a computer package for state space analysis is *SsfPack* of Koopman, Shephard and Doornik (1999, 2008) which we discuss next in more detail.

6.7.2 *SsfPack*

SsfPack is a suite of C routines for carrying out computations involving the statistical analysis of univariate and multivariate models in state space form. *SsfPack* allows for a range of different state space forms from simple time-invariant models to complicated time-varying models. Functions are specified which put standard models such as ARIMA and spline models into state space form. Routines are available for filtering, smoothing and simulation smoothing. Ready-to-use functions are provided for standard tasks such as likelihood evaluation, forecasting and signal extraction. The headers of these routines are documented by Koopman, Shephard and Doornik (1999); a more detailed documentation of the updated version 3.0 is presented in Koopman, Shephard and Doornik (2008). *SsfPack* can be used for implementing, fitting and analysing linear, Gaussian, nonlinear and/or non-Gaussian state space models relevant to many areas of time series analysis. A Gaussian illustration is given in Subsection 6.7.5.

6.7.3 The basic *SsfPack* functions

A list of *SsfPack* functions is given in Table 6.4. The functions are grouped into functions which put specific univariate models into state space form, functions which perform the basic filtering and smoothing operations and functions which execute specific important tasks for state space analysis such as likelihood evaluation. Table 6.4 consists of three columns : the first column contains the function name, the second column gives reference to the section number of the *SsfPack* documentation of Koopman, Shephard and Doornik (1999) and the third column describes the function with references to equation or section numbers in this book. This part of the package is freely available from `http://www.ssfpack.com`.

6.7.4 The extended *SsfPack* functions

In version 3.0 of *SsfPack* by Koopman, Shephard and Doornik (2008), further modifications of the computer routines are developed for the exact initial Kalman filter of Section 5.2, the exact initial state smoothing of Section 5.3, the exact initial disturbance smoothing of Section 5.4 and the computation of the diffuse likelihood functions defined in Section 7.2. The additional functions are listed in Table 6.5 in the same way as in Table 6.4.

Table 6.4 Functions in *SsfPack Basic* with section reference to documentation and with short description.

Models in state space form		
`AddSsfReg`	§3.3	adds regression effects (3.30) to state space.
`GetSsfArma`	§3.1	puts ARMA model (3.18) in state space.
`GetSsfReg`	§3.3	puts regression model (3.30) in state space.
`GetSsfSpline`	§3.4	puts cubic spline model (3.46) in state space.
`GetSsfStsm`	§3.2	puts structural time series model of §3.2 in state space.
`SsfCombine`	§6.2	combines system matrices of two models.
`SsfCombineSym`	§6.2	combines symmetric system matrices of two models.
General state space algorithms		
`KalmanFil`	§4.3	provides output of the Kalman filter in §4.3.2.
`KalmanSmo`	§4.4	provides output of the basic smoothing algorithm in §4.4.4.
`SimSmoDraw`	§4.5	provides a simulated sample.
`SimSmoWgt`	§4.5	provides output of the simulation smoother.
Ready-to-use functions		
`SsfCondDens`	§4.6	provides mean or a draw from the conditional density (4.81).
`SsfLik`	§5.1	provides loglikelihood function (7.4).
`SsfLikConc`	§5.1	provides concentrated loglikelihood function.
`SsfLikSco`	§5.1	provides score vector information (2.63).
`SsfMomentEst`	§5.2	provides output from prediction, forecasting and §5.3 smoothing.
`SsfRecursion`	§4.2	provides output of the state space recursion (3.1).

Table 6.5 Additional functions in *SsfPack Extended* with section reference to documentation and with short description.

General state space algorithms		
`KalmanInit`	§8.2.2	provides output of the exact initial
`KalmanFilEx`	§8.2.2	Kalman filter in §5.2.
`KalmanSmoEx`	§8.2.4	provides output of exact initial smoothing.
`KalmanFilSmoMeanEx`	§9.2	provides output to facilitate the augmented Kalman filter and smoother of §5.7.
Ready-to-use functions		
`SsfCondDensEx`	§9.4	provides the mean from the conditional density based on §§5.2 and 5.3.
`SsfLikEx`	§9.1.2	provides diffuse loglikelihood function (7.4).
`SsfLikConcEx`	§9.1.2	provides concentrated diffuse loglikelihood function.
`SsfLikScoEx`	§9.1.3	provides score vector information, see §7.3.3.
`SsfMomentEstEx`	§9.4	provides output from prediction, forecasting and based on §§5.2 and 5.3.
`SsfForecast`	§9.5	provides output of Kalman filter for forecasting, see §4.11.
`SsfWeights`	§9.6	provides output for filter and smoothing weights, see §4.8.

For all tasks involving the (exact initial) Kalman filter and smoother, the univariate treatment of multivariate series as described in Section 6.4 is implemented for the *SsfPack Extended* functions when the state space form represents a multivariate time series model. Furthermore, all computations involving unity and zero values are treated specifically such that they can handle sparse system matrices in a computationally efficient manner. These amendments lead to computationally faster and numerically more stable calculations. We will not discuss the *SsfPack* functions further but an example of *Ox* code, which utilises the link with the *SsfPack* library, is given in Subsection 6.7.5. This part of the package is a commercial product; more information can be obtained from `http://www.ssfpack.com`.

6.7.5 Illustration: spline smoothing

In the *Ox* code below we consider the continuous spline smoothing problem which aims at minimising (3.46) for a given value of λ. The aim of the program is to fit a spline through the Nile time series of Chapter 2 (see Subsection 2.2.5 for details). To illustrate that the *SsfPack* functions can deal with missing observations we have treated two parts of the data set as missing. The continuous spline model is easily put in state space form using the function `GetSsfSpline`. The smoothing parameter λ is chosen to take the value 2500 (the function requires the input of $\lambda^{-1} = 0.004$). We need to compute an estimator for the unknown scalar value of

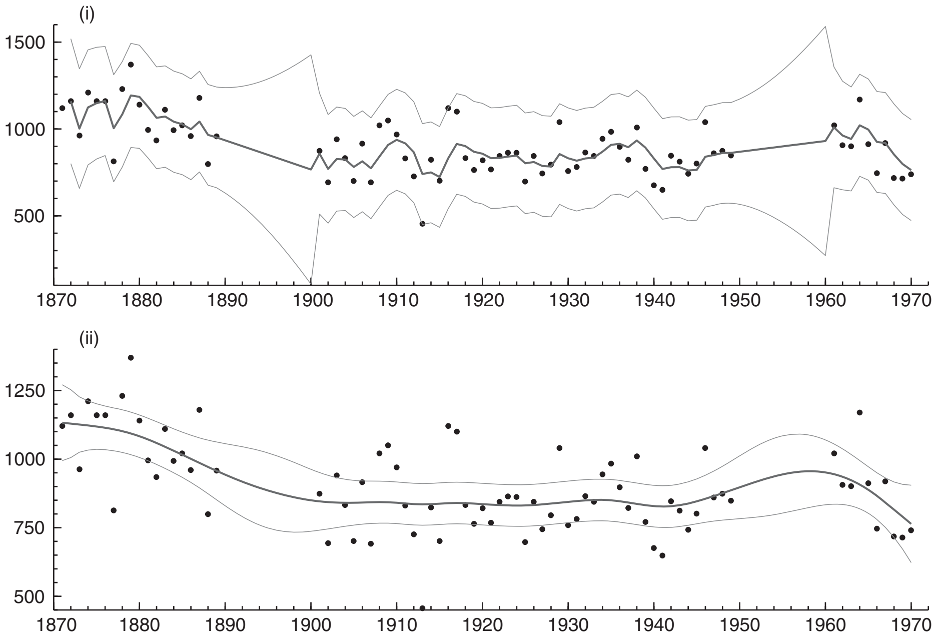

Fig. 6.1 Output of *Ox* program `spline.ox`: (i) Nile data (dots) with filtered estimate of spline function with 95% confidence interval; (ii) Nile data (dots) with smoothed estimate of spline function with 95% confidence interval.

σ_ζ^2 in (3.47) which can be obtained using the function `SsfLikEx`. After rescaling, the estimated spline function using filtering (`ST_FIL`) and smoothing (`ST_SMO`) is computed using the function `SsfMomentEst`. The output is presented in Fig. 6.1 and shows the filtered and smoothed estimates of the spline function. The two diagrams illustrate the point that filtering can be interpreted as extrapolation and that when observations are missing, smoothing is in fact interpolation.

```
 spline.ox
#include <oxstd.h>
#include <oxdraw.h>
#include <oxfloat.h>
#include <packages/ssfpack/ssfpack.h>
main()
{
  decl mdelta, mphi, momega, msigma, myt, mfil, msmo, cm, dlik, dvar;

  myt = loadmat("Nile.dat")';
  myt[][1890-1871:1900-1871] = M_NAN;  // set 1890..1900 to missing
  myt[][1950-1871:1960-1871] = M_NAN;  // set 1950..1960 to missing

  GetSsfSpline(0.004, <>, &mphi, &momega, &msigma); // SSF for spline
  SsfLikEx(&dlik, &dvar, myt, mphi, momega);          // need dVar
  cm = columns(mphi);                         // dimension of state
  momega *= dvar;                     // set correct scale of Omega
  SsfMomentEstEx(ST_FIL, &mfil,  myt, mphi, momega);
  SsfMomentEstEx(ST_SMO, &msmo, myt, mphi, momega);

  // note that first filtered estimator does not exist
  DrawTMatrix(0, myt, {"Nile"}, 1871, 1, 1);
  DrawTMatrix(0, mfil[cm][1:], {"Pred +/- 2SE"}, 1872, 1, 1, 0, 3);
  DrawZ(sqrt(mfil[2*cm+1][1:]), "", ZMODE_BAND, 2.0, 14);
  DrawTMatrix(1, myt, {"Nile"}, 1871, 1, 1);
  DrawTMatrix(1, msmo[cm][], {"Smooth +/- 2SE"}, 1871, 1, 1, 0, 3);
  DrawZ(sqrt(msmo[2*cm+1][]), "", ZMODE_BAND, 2.0, 14);
  ShowDrawWindow();
}
```

7 Maximum likelihood estimation of parameters

7.1 Introduction

So far we have developed methods for estimating parameters which can be placed in the state vector of model (4.12). In virtually all applications in practical work the models depend on additional parameters which have to be estimated from the data; for example, in the local level model (2.3) the variances σ_ε^2 and σ_η^2 are unknown and need to be estimated. In classical analyses, these additional parameters are assumed to be fixed but unknown whereas in Bayesian analyses they are assumed to be random variables. Because of the differences in assumptions the treatment of the two cases is not the same. In this chapter we deal with classical analyses in which the additional parameters are fixed and are estimated by maximum likelihood. The Bayesian treatment for these parameters is discussed as part of a general Bayesian discussion of state space models in Chapter 13 of Part II.

For the linear Gaussian model we shall show that the likelihood can be calculated by a routine application of the Kalman filter, even when the initial state vector is fully or partially diffuse. We also give the details of the computation of the likelihood when the univariate treatment of multivariate observations is adopted as suggested in Section 6.4. We go on to consider how the loglikelihood can be maximised by means of iterative numerical procedures. An important part in this process is played by the score vector and we show how this is calculated, both for the case where the initial state vector has a known distribution and for the diffuse case. A useful device for maximisation of the loglikelihood in some cases, particularly in the early stages of maximisation, is the EM algorithm; we give details of this for the linear Gaussian model. We go on to consider biases in estimates due to errors in parameter estimation. The chapter ends with a discussion of some questions of goodness-of-fit and diagnostic checks.

7.2 Likelihood evaluation

7.2.1 Loglikelihood when initial conditions are known

We first assume that the initial state vector has density $\mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known. The likelihood is

$$L(Y_n) = p(y_1, \dots, y_n) = p(y_1) \prod_{t=2}^{n} p(y_t|Y_{t-1}),$$

where $Y_t = (y_1', \dots, y_t')'$. In practice we generally work with the loglikelihood

$$\log L(Y_n) = \sum_{t=1}^{n} \log p(y_t|Y_{t-1}), \tag{7.1}$$

where $p(y_1|Y_0) = p(y_1)$. For model (3.1), $\mathrm{E}(y_t|Y_{t-1}) = Z_t a_t$. Putting $v_t = y_t - Z_t a_t$, $F_t = \mathrm{Var}(y_t|Y_{t-1})$ and substituting $\mathrm{N}(Z_t a_t, F_t)$ for $p(y_t|Y_{t-1})$ in (7.1), we obtain

$$\log L(Y_n) = -\frac{np}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{n} \left(\log|F_t| + v_t' F_t^{-1} v_t \right). \tag{7.2}$$

The quantities v_t and F_t are calculated routinely by the Kalman filter (4.24) so $\log L(Y_n)$ is easily computed from the Kalman filter output. We assume that F_t is nonsingular for $t = 1, \dots, n$. If this condition is not satisfied initially it is usually possible to redefine the model so that it is satisfied. The representation (7.2) of the loglikelihood was first given by Schweppe (1965). Harvey (1989, §3.4) refers to it as the *prediction error decomposition.*

7.2.2 Diffuse loglikelihood

We now consider the case where some elements of α_1 are diffuse. As in Section 5.1, we assume that $\alpha_1 = a + A\delta + R_0\eta_0$ where a is a known constant vector, $\delta \sim \mathrm{N}(0, \kappa I_q)$, $\eta_0 \sim \mathrm{N}(0, Q_0)$ and $A'R_0 = 0$, giving $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ where $P_1 = \kappa P_\infty + P_*$ and $\kappa \to \infty$. From (5.6) and (5.7),

$$F_t = \kappa F_{\infty,t} + F_{*,t} + O(\kappa^{-1}) \quad \text{with} \quad F_{\infty,t} = Z_t P_{\infty,t} Z_t', \tag{7.3}$$

where by definition of d, $P_{\infty,t} \neq 0$ for $t = 1, \dots, d$. The number of diffuse elements in α_1 is q which is the dimensionality of vector δ. Thus the loglikelihood (7.2) will contain a term $-\frac{1}{2} q \log 2\pi\kappa$ so $\log L(Y_n)$ will not converge as $\kappa \to \infty$. Following de Jong (1991), we therefore define the *diffuse loglikelihood* as

$$\log L_d(Y_n) = \lim_{\kappa \to \infty} \left[\log L(Y_n) + \frac{q}{2} \log \kappa \right]$$

and we work with $\log L_d(Y_n)$ in place of $\log L(Y_n)$ for estimation of unknown parameters in the diffuse case. Similar definitions for the diffuse loglikelihood function have been adopted by Harvey and Phillips (1979) and Ansley and Kohn (1986). As in Section 5.2, and for the same reasons, we assume that $F_{\infty,t}$ is positive definite or is a zero matrix. We also assume that q is a multiple of p. This covers the important special case of univariate series and is generally satisfied in practice for multivariate series; if not, the series can be dealt with as if it were univariate as in Section 6.4.

Suppose first that $F_{\infty,t}$ is positive definite and therefore has rank p. From (7.3) we have for $t = 1, \dots, d$,

$$F_t^{-1} = \kappa^{-1} F_{\infty,t}^{-1} + O(\kappa^{-2}).$$

It follows that

$$\begin{aligned} -\log|F_t| &= \log\left|F_t^{-1}\right| = \log\left|\kappa^{-1} F_{\infty,t}^{-1} + O(\kappa^{-2})\right| \\ &= -p \log \kappa + \log\left|F_{\infty,t}^{-1} + O(\kappa^{-1})\right|, \end{aligned}$$

and

$$\lim_{\kappa\to\infty} (-\log|F_t| + p \log \kappa) = \log\left|F_{\infty,t}^{-1}\right| = -\log|F_{\infty,t}|.$$

Moreover,

$$\begin{aligned} \lim_{\kappa\to\infty} v_t' F_t^{-1} v_t &= \lim_{\kappa\to\infty} \left[v_t^{(0)} + \kappa^{-1} v_t^{(1)} + O(\kappa^{-2})\right]' \left[\kappa^{-1} F_{\infty,t}^{-1} + O(\kappa^{-2})\right] \\ &\qquad \times \left[v_t^{(0)} + \kappa^{-1} v_t^{(1)} + O(\kappa^{-2})\right] \\ &= 0 \end{aligned}$$

for $t = 1, \dots, d$, where $v_t^{(0)}$ and $v_t^{(1)}$ are defined in Subsection 5.2.1.

When $F_{\infty,t} = 0$, it follows from Subsection 5.2.1 that $F_t = F_{*,t} + O(\kappa^{-1})$ and $F_t^{-1} = F_{*,t}^{-1} + O(\kappa^{-1})$. Consequently,

$$\lim_{\kappa\to\infty} (-\log|F_t|) = -\log|F_{*,t}| \quad \text{and} \quad \lim_{\kappa\to\infty} v_t' F_t^{-1} v_t = v_t^{(0)\prime} F_{*,t}^{-1} v_t^{(0)}.$$

Putting these results together, we obtain the diffuse loglikelihood as

$$\log L_d(Y_n) = -\frac{np}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{d} w_t - \frac{1}{2} \sum_{t=d+1}^{n} \left(\log|F_t| + v_t' F_t^{-1} v_t\right), \quad (7.4)$$

where

$$w_t = \begin{cases} \log|F_{\infty,t}|, & \text{if } F_{\infty,t} \text{ is positive definite}, \\ \log|F_{*,t}| + v_t^{(0)\prime} F_{*,t}^{-1} v_t^{(0)}, & \text{if } F_{\infty,t} = 0, \end{cases}$$

for $t = 1, \dots, d$. The expression (7.4) for the diffuse loglikelihood is given by Koopman (1997).

7.2.3 Diffuse loglikelihood via augmented Kalman filter

In the notation of Subsection 5.7.3, the joint density of δ and Y_n for given κ is

$$\begin{aligned} p(\delta, Y_n) &= p(\delta)p(Y_n|\delta) \\ &= p(\delta)\sum_{t=1}^{n} p(v_{\delta,t}) \\ &= (2\pi)^{-(np+q)/2}\kappa^{-q/2}\prod_{t=1}^{n}|F_{\delta,t}|^{-1/2} \\ &\quad \times \exp\left[-\frac{1}{2}\left(\frac{\delta'\delta}{\kappa} + S_{a,n} + 2b_n'\delta + \delta' S_{A,n}\delta\right)\right], \end{aligned} \tag{7.5}$$

where $v_{\delta,t}$ is defined in (5.35), b_n and $S_{A,n}$ are defined in (5.40) and $S_{a,n} = \sum_{t=1}^{n} v_{a,t}' F_{\delta,t}^{-1} v_{a,t}$. From (5.41) we have $\bar{\delta}_n = \mathrm{E}(\delta|Y_n) = -(S_{A,n} + \kappa^{-1}I_q)^{-1}b_n$. The exponent of (7.5) can now be rewritten as

$$-\frac{1}{2}[S_{a,n} + (\delta - \bar{\delta}_n)'(S_{A,n} + \kappa^{-1}I_q)(\delta - \bar{\delta}_n) - \bar{\delta}_n'(S_{A,n} + \kappa^{-1}I_q)\bar{\delta}_n],$$

as is easily verified. Integrating out δ from $p(\delta, Y_n)$ we obtain the marginal density of Y_n. After taking logs, the loglikelihood appears as

$$\begin{aligned} \log L(Y_n) &= -\frac{np}{2}\log 2\pi - \frac{q}{2}\log\kappa - \frac{1}{2}\log|S_{A,n} + \kappa^{-1}I_q| \\ &\quad - \frac{1}{2}\sum_{t=1}^{n}\log|F_{\delta,t}| - \frac{1}{2}[S_{a,n} - \bar{\delta}_n'(S_{A,n} + \kappa^{-1}I_q)\bar{\delta}_n]. \end{aligned} \tag{7.6}$$

Adding $\frac{q}{2}\log\kappa$ and letting $\kappa \to \infty$ we obtain the diffuse loglikelihood

$$\begin{aligned} \log L_d(Y_n) &= -\frac{np}{2}\log 2\pi - \frac{1}{2}\log|S_{A,n}| - \frac{1}{2}\sum_{t=1}^{n}\log|F_{\delta,t}| \\ &\quad - \frac{1}{2}\left(S_{a,n} - b_n' S_{A,n}^{-1} b_n\right), \end{aligned} \tag{7.7}$$

which is due to de Jong (1991). In spite of its very different structure (7.7) necessarily has the same numerical value as (7.4).

It is shown in Subsection 5.7.3 that the augmented Kalman filter can be collapsed at time point $t = d$. We could therefore form a partial likelihood based on Y_d for fixed κ, integrate out δ and let $\kappa \to \infty$ as in (7.7). Subsequently we could add the contribution from innovations $v_{d+1}, \dots, v_n$ obtained from the collapsed Kalman filter. However, we will not give detailed formulae here.

These results were originally derived by de Jong (1988b, 1999). The calculations required to compute (7.7) are more complicated than those required to compute (7.4). This is another reason why we ourselves prefer the initialisation technique of Section 5.2 to the augmentation device of Section 5.7. A further reason for prefering our computation of (7.4) is given in Subsection 7.3.5.

7.2.4 Likelihood when elements of initial state vector are fixed but unknown

Now let us consider the case where δ is treated as fixed. The density of Y_n given δ is, as in the previous section,

$$p(Y_n|\delta) = (2\pi)^{-np/2} \prod_{t=1}^{n} |F_{\delta,t}|^{-1/2} \exp\left[-\frac{1}{2}(S_{a,n} + 2b_n'\delta_n + \delta_n' S_{A,n}\delta_n)\right]. \qquad (7.8)$$

The usual way to remove the influence of an unknown parameter vector such as δ from the likelihood is to estimate it by its maximum likelihood estimate, $\hat{\delta}_n$ in this case, and to employ the concentrated loglikelihood $\log L_c(Y_n)$ obtained by substituting $\hat{\delta}_n = -S_{A,n}^{-1}b_n$ for δ in $p(Y_n|\delta)$. This gives

$$\log L_c(Y_n) = -\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}|F_{\delta,t}| - \frac{1}{2}(S_{a,n} - b_n' S_{A,n}^{-1} b_n). \qquad (7.9)$$

Comparing (7.7) and (7.9) we see that the only difference between them is the presence in (7.7) of the term $-\frac{1}{2}\log|S_{A,n}|$. The relation between (7.7) and (7.9) was demonstrated by de Jong (1988b) using a different argument. A modification of the augmented Kalman filter and the corresponding diffuse loglikelihood function is proposed by Francke, Koopman and de Vos (2010). Their modification ensures that the loglikelihood function has the same value regardless of the way the regression effects are treated in the model.

Harvey and Shephard (1990) argue that parameter estimation should preferably be based on the loglikelihood function (7.4) for which the initial vector δ is treated as diffuse not fixed. They have shown for the local level model of Chapter 2 that maximising (7.9) with respect to the signal to noise ratio q leads to a much higher probability of estimating q to be zero compared to maximising (7.4). This is undesirable from a forecasting point of view since this results in no discounting of past observations.

7.2.5 Likelihood when a univariate treatment of multivariate series is employed

When the univariate treatment of the Kalman filter for multivariate time series is considered as in Section 6.4, we transform the vector time series $y_1, \dots, y_n$ into the univariate time series

$$y_{1,1}, \dots, y_{1,p_1}, y_{2,1}, \dots, y_{n,p_n}.$$

The univariate Kalman filter (6.12) and (6.13) can then be applied to the model equations (6.10) and (6.11) for this univariate series. It produces the prediction error $v_{t,i}$ and its scalar variance $F_{t,i}$ for $t = 1, \dots, n$ and $i = 1, \dots, p_t$, where the prediction error is the error in predicting $y_{t,i}$ as a function of the 'past' observations $y_{1,1}, \dots, y_{1,p_1}, y_{2,1}, \dots, y_{t,i-1}$ for $i = 2, \dots p_t$ and $y_{1,1}, \dots, y_{1,p_1}, y_{2,1}, \dots, y_{t-1,p_{t-1}}$ for $i = 1$. Since the model (6.10) and (6.11) for the univariate series is fully consistent with the original model for the vector time series $y_1, \dots, y_n$ and the Kalman filter is adapted correctly, the loglikelihood function when initial conditions are known is given by

$$\log L(Y_n) = -\frac{1}{2} \sum_{t=1}^{n} \left[p_t^* \log 2\pi + \sum_{i=1}^{p_t} \iota_{t,i} (\log F_{t,i} + v_{t,i}^2 \,/\, F_{t,i}) \right].$$

where $\iota_{t,i}$ equals zero if $F_{t,i} = 0$ and unity otherwise, and $p_t^* = \sum_{i=1}^{p_t} \iota_{t,i}$. The occurence of $F_{t,i} = 0$ is due to the singularity of F_t as discussed in Section 6.4.

The variables associated with the exact initial Kalman filter of Section 5.2 can also redefined when it is applied to the transformation of a multivariate series into a univariate series. When the univariate treatment is used, the relevant variables for the diffuse loglikelihood function considered in Subsection 7.2.2 are given by the scalars $v_{t,i}^{(0)}$, $F_{\infty,t,i}$ and $F_{*,t,i}$ for $t = 1, \dots, d$, and $v_{t,i}$ and $F_{t,i}$ for $t = d + 1, \dots, n$, with $i = 1, \dots, p_t$. The diffuse loglikelihood function (7.4) is then computed by

$$\begin{aligned} \log L_d(Y_n) = &- \frac{1}{2} \sum_{t=1}^{n} \sum_{i=1}^{p_t} \iota_{t,i} \log 2\pi - \frac{1}{2} \sum_{t=1}^{d} \sum_{i=1}^{p_t} w_{t,i} \\ &- \frac{1}{2} \sum_{t=d+1}^{n} \sum_{i=1}^{p_t} \iota_{t,i} (\log F_{t,i} + v_{t,i}^2 \,/\, F_{t,i}), \end{aligned}$$

where $\iota_{t,i}$ equal zero if $F_{*,t,i} = 0$ and unity otherwise, and where

$$w_{t,i} = \begin{cases} \log F_{\infty,t,i}, & \text{if} \quad F_{\infty,t,i} > 0, \\ \iota_{t,i} (\log F_{*,t,i} + v_t^{(0)\,2} \,/\, F_{*,t,i}), & \text{if} \quad F_{\infty,t,i} = 0, \end{cases}$$

for $t = 1, \dots, d$ and $i = 1, \dots, p_t$.

7.2.6 Likelihood when the model contains regression effects

When regression effects are present in the state space model, similar adjustments as for the diffuse loglikelihood function are required. The diffuse loglikelihood can be regarded as a loglikelihood function of a linear transformation of the original time series. When such a loglikelihood function is used for parameter estimation, the transformation must not depend on the parameter vector itself. When it does, the likelihood function is not a smooth function with respect to the parameter vector. Francke, Koopman and de Vos (2010) show that a proper transformation can be obtained via a simple modification of the augmented Kalman filter. They provide the details with a comparison of the different likelihood functions and some illustrations.

7.2.7 Likelihood when large observation vector is collapsed

When a large $p \times 1$ observation vector y_t needs to be treated in the analysis, we proposed in Section 6.5 to collapse y_t into a transformed observation vector with its dimension equal to the dimension of the state vector m. It is shown in Section 6.5 that the collapse strategy becomes computationally beneficial when $p >> m$. The Kalman filter can be applied to a small observation vector without additional costs. We show next that the likelihood function for the original model can be computed using the Kalman filter for a small observation vector together with some additional computing. Consider y_t^* and y_t^+ as defined in Subsection 6.5.2, or their counterparts $\bar{y}_t^*$ and $\bar{y}_t^+$ as defined in Subsection 6.5.3, for which the models

$$y_t^* = \alpha_t + \varepsilon_t^*, \qquad y_t^+ = \varepsilon_t^+,$$

are adopted where $\varepsilon_t^* \sim \mathrm{N}(0, H_t^*)$ and $\varepsilon_t^+ \sim \mathrm{N}(0, H_t^+)$ are serially and mutually independent. It follows that the original loglikelihood function is subject to the relation

$$\log L(Y_n) = \log L(Y_n^*) + \log L(Y_n^+) + \sum_{t=1}^{n} \log |A_t|,$$

where $Y_n^* = (y_1^{*\prime}, \ldots, y_n^{*\prime})'$, $Y_n^+ = (y_1^{+\prime}, \ldots, y_n^{+\prime})'$ and $A_t = (A_t^{*\prime}, A_t^{+\prime})'$ such that $A_t y_t = (y_t^{*\prime}, y_t^{+\prime})'$ for $t = 1, \ldots, n$. The term $|A_t|$ can be expressed more conveniently via the relations

$$|A_t|^2 = |A_t A_t'| = |H_t^{-1}||A_t H_t A_t'| = |H_t|^{-1}|H_t^*||H_t^+|.$$

Since the scaling of matrix A_t^+ is not relevant for its construction in Subsections 6.5.2 and 6.5.3, we choose A_t^+ such that $|H_t^+| = 1$. We then have

$$\log |A_t| = \frac{1}{2} \log \frac{|H_t^*|}{|H_t|}, \qquad t = 1, \ldots, n,$$

which can be computed efficiently since matrix H_t^* has a small dimension while H_t is the variance matrix of the original observation disturbance ε_t and is typically a diagonal matrix or has a convenient structure. The loglikelihood function $\log L(Y_n^*)$ is computed by the Kalman filter applied to the model $y_t^* = \alpha_t + \varepsilon_t^*$ and $\log L(Y_n^+)$ is given by

$$\log L(Y_n^+) = -\frac{(p-m)n}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{n} y_t^{+\prime} (H_t^+)^{-1} y_t^+,$$

since we have $|H_t^+| = 1$. The term $y_t^{+\prime}(H_t^+)^{-1} y_t^+$ insists that matrix A_t^+ actually needs to be computed. However, it is shown in Jungbacker and Koopman (2008, Lemma 2) that this term can be computed as

$$y_t^{+\prime}(H_t^+)^{-1} y_t^+ = e_t' H_t^{-1} e_t,$$

where $e_t = y_t - Z_t y_t^*$ in or, more generally, $e_t = y_t - Z_t \bar{y}_t^*$ for any nonsingular matrix C_t as defined in Subsection 6.5.3. The computation of $e_t' H_t^{-1} e_t$ is simple and does not require the construction of matrix A_t^+ or $\bar{A}_t^+$.

7.3 Parameter estimation

7.3.1 Introduction

So far in this book we have assumed that the system matrices Z_t, H_t, T_t, R_t and Q_t in model (3.1) are known for $t = 1, \ldots, n$. We now consider the more usual situation in which at least some of the elements of these matrices depend on a vector ψ of unknown parameters. We shall estimate ψ by maximum likelihood. To make explicit the dependence of the loglikelihood on ψ we write $\log L(Y_n|\psi)$, $\log L_d(Y_n|\psi)$ and $\log L_c(Y_n|\psi)$. In the diffuse case we shall take it for granted that for models of interest, estimates of ψ obtained by maximising $\log L(Y_n|\psi)$ for fixed κ converge to the estimates obtained by maximising the diffuse loglikelihood $\log L_d(Y_n|\psi)$ as $\kappa \to \infty$.

7.3.2 Numerical maximisation algorithms

A wide range of numerical search algorithms are available for maximising the loglikelihood with respect to unknown parameters. Many of these are based on Newton's method which solves the equation

$$\partial_1(\psi) = \frac{\partial \log L(Y_n|\psi)}{\partial \psi} = 0, \tag{7.10}$$

using the first-order Taylor series

$$\partial_1(\psi) \simeq \tilde{\partial}_1(\psi) + \tilde{\partial}_2(\psi)(\psi - \tilde{\psi}), \tag{7.11}$$

for some trial value $\tilde{\psi}$, where

$$\tilde{\partial}_1(\psi) = \partial_1(\psi)|_{\psi=\tilde{\psi}}, \qquad \tilde{\partial}_2(\psi) = \partial_2(\psi)|_{\psi=\tilde{\psi}},$$

with

$$\partial_2(\psi) = \frac{\partial^2 \log L(Y_n|\psi)}{\partial \psi \partial \psi'}. \tag{7.12}$$

By equating (7.11) to zero we obtain a revised value $\bar{\psi}$ from the expression

$$\bar{\psi} = \tilde{\psi} - \tilde{\partial}_2(\psi)^{-1} \tilde{\partial}_1(\psi).$$

This process is repeated until it converges or until a switch is made to another optimisation method. If the Hessian matrix $\partial_2(\psi)$ is negative definite for all ψ the loglikelihood is said to be concave and a unique maximum exists for the likelihood. The *gradient* $\partial_1(\psi)$ determines the direction of the step taken to the optimum and the Hessian modifies the size of the step. It is possible to overstep the maximum in the direction determined by the vector

$$\tilde{\pi}(\psi) = -\tilde{\partial}_2(\psi)^{-1}\tilde{\partial}_1(\psi),$$

and therefore it is common practice to include a line search along the gradient vector within the optimisation process. We obtain the algorithm

$$\bar{\psi} = \tilde{\psi} + s\tilde{\pi}(\psi),$$

where various methods are available to find the optimum value for s which is usually found to be between 0 and 1. In practice it is often computationally demanding or impossible to compute $\partial_1(\psi)$ and $\partial_2(\psi)$ analytically. Numerical evaluation of $\partial_1(\psi)$ is usually feasible. A variety of computational devices are available to approximate $\partial_2(\psi)$ in order to avoid computing it analytically or numerically. For example, the *STAMP* package of Koopman, Harvey, Doornik and Shephard (2010) and the *Ox* matrix programming system of Doornik (2010) both use the so-called BFGS (Broyden–Fletcher–Goldfarb–Shannon) method which approximates the Hessian matrix using a device in which at each new value for ψ a new approximate inverse Hessian matrix is obtained via the recursion

$$\bar{\partial}_2(\psi)^{-1} = \tilde{\partial}_2(\psi)^{-1} + \left(s + \frac{g'g^*}{\tilde{\pi}(\psi)'g}\right)\frac{\tilde{\pi}(\psi)\tilde{\pi}(\psi)'}{\tilde{\pi}(\psi)'g} - \frac{\tilde{\pi}(\psi)g^{*\prime} + g^*\tilde{\pi}(\psi)'}{\tilde{\pi}(\psi)'g},$$

where g is the difference between the gradient $\tilde{\partial}_1(\psi)$ and the gradient for a trial value of ψ prior to $\tilde{\psi}$ and

$$g^* = \tilde{\partial}_2(\psi)^{-1}g.$$

The BFGS method ensures that the approximate Hessian matrix remains negative definite. The details and derivations of the Newton's method of optimisation and the BFGS method in particular can be found, for example, in Fletcher (1987).

Model parameters are sometimes constrained. For example, the parameters in the local level model (2.3) must satisfy the constraints $\sigma_\varepsilon^2 \geq 0$ and $\sigma_\eta^2 \geq 0$ with $\sigma_\varepsilon^2 + \sigma_\eta^2 > 0$. However, the introduction of constraints such as these within the numerical procedure is inconvenient and it is preferable that the maximisation is performed with respect to quantities which are unconstrained. For this example we therefore make the transformations $\psi_\varepsilon = \frac{1}{2}\log\sigma_\varepsilon^2$ and $\psi_\eta = \frac{1}{2}\log\sigma_\eta^2$ where $-\infty < \psi_\varepsilon, \psi_\eta < \infty$, thus converting the problem to one in unconstrained maximisation. The parameter vector is $\psi = [\psi_\varepsilon, \psi_\eta]'$. Similarly, if we have a parameter χ which is restricted to the range $[-a, a]$ where a is positive we can make a transformation to ψ_χ for which

$$\chi = \frac{a\psi_\chi}{\sqrt{1+\psi_\chi^2}}, \qquad -\infty < \psi_\chi < \infty.$$

7.3.3 The score vector

We now consider details of the calculation of the gradient or *score vector*

$$\partial_1(\psi) = \frac{\partial \log L(Y_n|\psi)}{\partial \psi}.$$

As indicated in the last section, this vector is important in numerical maximisation since it specifies the direction in the parameter space along which a search should be made.

We begin with the case where the initial vector α_1 has the distribution $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ where a_1 and P_1 are known. Let $p(\alpha, Y_n|\psi)$ be the joint density of α and Y_n, let $p(\alpha|Y_n, \psi)$ be the conditional density of α given Y_n and let $p(Y_n|\psi)$ be the marginal density of Y_n for given ψ. We now evaluate the score vector $\partial \log L(Y_n|\psi)/\partial\psi = \partial \log p(Y_n|\psi)/\partial\psi$ at the trial value $\tilde{\psi}$. We have

$$\log p(Y_n|\psi) = \log p(\alpha, Y_n|\psi) - \log p(\alpha|Y_n, \psi).$$

Let $\tilde{\mathrm{E}}$ denote expectation with respect to density $p(\alpha|Y_n, \tilde{\psi})$. Since $p(Y_n|\psi)$ does not depend on α, taking $\tilde{\mathrm{E}}$ of both sides gives

$$\log p(Y_n|\psi) = \tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)] - \tilde{\mathrm{E}}[\log p(\alpha|Y_n, \psi)].$$

To obtain the score vector at $\tilde{\psi}$, we differentiate both sides with respect to ψ and put $\psi = \tilde{\psi}$. Assuming that differentiating under integral signs is legitimate,

$$\begin{aligned}\tilde{\mathrm{E}}\left[\left.\frac{\partial \log p(\alpha|Y_n, \psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}}\right] &= \int \frac{1}{p(\alpha|Y_n, \tilde{\psi})} \left.\frac{\partial p(\alpha|Y_n, \psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} p(\alpha|Y_n, \tilde{\psi})\, d\alpha \\ &= \left.\frac{\partial}{\partial \psi} \int p(\alpha|Y_n, \psi)\, d\alpha\right|_{\psi=\tilde{\psi}} = 0.\end{aligned}$$

Thus

$$\left.\frac{\partial \log p(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} = \tilde{\mathrm{E}}\left[\frac{\partial \log p(\alpha, Y_n|\psi)}{\partial \psi}\right]\Bigg|_{\psi=\tilde{\psi}}.$$

With substitutions $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$ and $\varepsilon_t = y_t - Z_t\alpha_t$ and putting $\alpha_1 - a_1 = \eta_0$ and $P_1 = Q_0$, we obtain

$$\begin{aligned}\log p(\alpha, Y_n|\psi) &= \text{constant} \\ &\quad - \frac{1}{2}\sum_{t=1}^{n}\left(\log|H_t| + \log|Q_{t-1}| + \varepsilon_t' H_t^{-1}\varepsilon_t + \eta_{t-1}' Q_{t-1}^{-1}\eta_{t-1}\right). \qquad (7.13)\end{aligned}$$

On taking the expectation $\tilde{\mathrm{E}}$ and differentiating with respect to ψ, this gives the score vector at $\psi = \tilde{\psi}$,

$$\left.\frac{\partial \log L(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} = -\frac{1}{2}\frac{\partial}{\partial \psi}\sum_{t=1}^{n}\Big[\big(\log|H_t| + \log|Q_{t-1}| \\ + \mathrm{tr}\big[\{\hat{\varepsilon}_t\hat{\varepsilon}_t' + \mathrm{Var}(\varepsilon_t|Y_n)\}H_t^{-1}\big] \\ + \mathrm{tr}\big[\{\hat{\eta}_{t-1}\hat{\eta}_{t-1}' + \mathrm{Var}(\eta_{t-1}|Y_n)\}Q_{t-1}^{-1}\big]\big|\psi\big)\big|_{\psi=\tilde{\psi}}\Big], \tag{7.14}$$

where $\hat{\varepsilon}_t$, $\hat{\eta}_{t-1}$, $\mathrm{Var}(\varepsilon_t|Y_n)$ and $\mathrm{Var}(\eta_{t-1}|Y_n)$ are obtained for $\psi = \tilde{\psi}$ as in Section 4.5.

Only the terms in H_t and Q_t in (7.14) require differentiation with respect to ψ. Since in practice H_t and Q_t are often simple functions of ψ, this means that the score vector is often easy to calculate, which can be a considerable advantage in numerical maximisation of the loglikelihood. A similar technique can be developed for the system matrices Z_t and T_t but this requires more computations which involve the state smoothing recursions. Koopman and Shephard (1992), to whom the result (7.14) is due, therefore conclude that the score values for ψ associated with system matrices Z_t and T_t can be evaluated better numerically than analytically.

We now consider the diffuse case. In Section 5.1 we specified the initial state vector α_1 as

$$\alpha_1 = a + A\delta + R_0\eta_0, \qquad \delta \sim \mathrm{N}(0, \kappa I_q), \qquad \eta_0 \sim \mathrm{N}(0, Q_0),$$

where Q_0 is nonsingular. Equation (7.13) is still valid except that $\alpha_1 - a_1 = \eta_0$ is now replaced by $\alpha_1 - a = A\delta + R_0\eta_0$ and $P_1 = \kappa P_\infty + P_*$ where $P_* = R_0 Q_0 R_0'$. Thus for finite κ the term

$$-\frac{1}{2}\frac{\partial}{\partial \psi}(q \log \kappa + \kappa^{-1}\mathrm{tr}\{\hat{\delta}\hat{\delta}' + \mathrm{Var}(\delta|Y_n)\})$$

must be included in (7.14). Defining

$$\left.\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} = \lim_{\kappa\to\infty}\frac{\partial}{\partial \psi}\left[\log L(Y_n|\psi) + \frac{q}{2}\log \kappa\right],$$

analogously to the definition of $\log L_d(Y_n)$ in Subsection 7.2.2, and letting $\kappa \to \infty$, we have that

$$\left.\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} = \left.\frac{\partial \log L(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}}, \tag{7.15}$$

which is given in (7.14). In the event that α_1 consists only of diffuse elements, so the vector η_0 is null, the terms in Q_0 disappear from (7.14).

As an example consider the local level model (2.3) with η replaced by ξ, for which

$$\psi = \begin{pmatrix} \psi_\varepsilon \\ \psi_\xi \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\log\sigma_\varepsilon^2 \\ \frac{1}{2}\log\sigma_\xi^2 \end{pmatrix},$$

with a diffuse initialisation for α_1. Then ψ is the unknown parameter vector of the kind mentioned in Section 7.1. We have on substituting $y_t - \alpha_t = \varepsilon_t$ and $\alpha_{t+1} - \alpha_t = \xi_t$,

$$\begin{aligned}\log p(\alpha, Y_n|\psi) = &-\frac{2n-1}{2}\log 2\pi - \frac{n}{2}\log\sigma_\varepsilon^2 - \frac{n-1}{2}\log\sigma_\xi^2 \\ &- \frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^n \varepsilon_t^2 - \frac{1}{2\sigma_\xi^2}\sum_{t=2}^n \xi_{t-1}^2,\end{aligned}$$

and

$$\begin{aligned}\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)] = &-\frac{2n-1}{2}\log 2\pi - \frac{n}{2}\log\sigma_\varepsilon^2 - \frac{n-1}{2}\log\sigma_\xi^2 - \frac{1}{2\sigma_\varepsilon^2} \\ &\times \sum_{t=1}^n \left\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\right\} - \frac{1}{2\sigma_\xi^2}\sum_{t=2}^n \left\{\hat{\xi}_{t-1}^2 + \mathrm{Var}(\xi_{t-1}|Y_n)\right\},\end{aligned}$$

where the conditional means and variances for ε_t and ξ_t are obtained from the Kalman filter and disturbance smoother with σ_ε^2 and σ_ξ^2 implied by $\psi = \tilde{\psi}$. To obtain the score vector, we differentiate both sides with respect to ψ where we note that, with $\psi_\varepsilon = \frac{1}{2}\log\sigma_\varepsilon^2$,

$$\begin{aligned}\frac{\partial}{\partial\sigma_\varepsilon^2}\left[\log\sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2}\left\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\right\}\right] &= \frac{1}{\sigma_\varepsilon^2} - \frac{1}{\sigma_\varepsilon^4}\left\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\right\}, \\ \frac{\partial\sigma_\varepsilon^2}{\partial\psi_\varepsilon} &= 2\sigma_\varepsilon^2.\end{aligned}$$

The terms $\hat{\varepsilon}_t$ and $\mathrm{Var}(\varepsilon_t|Y_n)$ do not vary with ψ since they have been calculated on the assumption that $\psi = \tilde{\psi}$. We obtain

$$\begin{aligned}\frac{\partial \log L_d(Y_n|\psi)}{\partial\psi_\varepsilon} &= -\frac{1}{2}\frac{\partial}{\partial\psi_\varepsilon}\sum_{t=1}^n\left[\log\sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2}\left\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\right\}\right] \\ &= -n + \frac{1}{\sigma_\varepsilon^2}\sum_{t=1}^n\left\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\right\}.\end{aligned}$$

In a similar way we have

$$\begin{aligned}\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi_\xi} &= -\frac{1}{2}\frac{\partial}{\partial \psi_\xi}\sum_{t=2}^{n}\left[\log \sigma_\xi^2 + \frac{1}{\sigma_\xi^2}\left\{\hat{\xi}_{t-1}^2 + \mathrm{Var}(\xi_{t-1}|Y_n)\right\}\right] \\ &= 1 - n + \frac{1}{\sigma_\xi^2}\sum_{t=2}^{n}\left\{\hat{\xi}_{t-1}^2 + \mathrm{Var}(\xi_{t-1}|Y_n)\right\}.\end{aligned}$$

The score vector for ψ of the local level model evaluated at $\psi = \tilde{\psi}$ is therefore

$$\left.\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} = \left[\begin{array}{c} \tilde{\sigma}_\varepsilon^2 \sum_{t=1}^{n}\left(u_t^2 - D_t\right) \\ \tilde{\sigma}_\xi^2 \sum_{t=2}^{n}\left(r_{t-1}^2 - N_{t-1}\right)\end{array}\right],$$

with $\tilde{\sigma}_\varepsilon^2$ and $\tilde{\sigma}_\xi^2$ from $\tilde{\psi}$. This result follows since from Subsections 2.5.1 and 2.5.2 $\hat{\varepsilon}_t = \tilde{\sigma}_\varepsilon^2 u_t$, $\mathrm{Var}(\varepsilon_t|Y_n) = \tilde{\sigma}_\varepsilon^2 - \tilde{\sigma}_\varepsilon^4 D_t$, $\hat{\xi}_t = \tilde{\sigma}_\xi^2 r_t$ and $\mathrm{Var}(\xi_t|Y_n) = \tilde{\sigma}_\xi^2 - \tilde{\sigma}_\xi^4 N_t$.

It is very satisfactory that after so much algebra we obtain such a simple expression for the score vector which can be computed efficiently using the disturbance smoothing equations of Section 4.5. We can compute the score vector for the diffuse case efficiently because it is shown in Section 5.4 that no extra computing is required for disturbance smoothing when dealing with a diffuse initial state vector. Finally, score vector elements associated with variances or variance matrices in more complicated models such as multivariate structural time series models continue to have similar relatively simple expressions. Koopman and Shephard (1992) give for these models the score vector for parameters in H_t, R_t and Q_t as the expression

$$\begin{aligned}\left.\frac{\partial \log L_d(Y_n|\psi)}{\partial \psi}\right|_{\psi=\tilde{\psi}} &= \frac{1}{2}\sum_{t=1}^{n}\mathrm{tr}\left\{(u_t u_t' - D_t)\frac{\partial H_t}{\partial \psi}\right\} \\ &\quad + \frac{1}{2}\sum_{t=2}^{n}\mathrm{tr}\left\{(r_{t-1}r_{t-1}' - N_{t-1})\frac{\partial R_t Q_t R_t'}{\partial \psi}\right\}\right|_{\psi=\tilde{\psi}}, \qquad (7.16)\end{aligned}$$

where u_t, D_t, r_t and N_t are evaluated by the Kalman filter and smoother as discussed in Sections 4.5 and 5.4.

7.3.4 The EM algorithm

The EM algorithm is a well-known tool for iterative maximum likelihood estimation which for many state space models has a particularly neat form. The earlier EM methods for the state space model were developed by Shumway and Stoffer (1982) and Watson and Engle (1983). The EM algorithm can be used either entirely instead of, or in place of the early stages of, direct numerical maximisation of the loglikelihood. It consists of an E-step (expectation) and

M-step (maximisation) for which the former involves the evaluation of the conditional expectation $\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)]$ and the latter maximises this expectation with respect to the elements of ψ. The details of estimating unknown elements in H_t and Q_t are given by Koopman (1993) and they are close to those required for the evaluation of the score function. Taking first the case of a_1 and P_1 known and starting with (7.13), we evaluate $\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)]$ and as in (7.14) we obtain

$$\begin{aligned}\frac{\partial}{\partial\psi}\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)] = -\frac{1}{2}\frac{\partial}{\partial\psi}\sum_{t=1}^{n}\big[&\log|H_t| + \log|Q_{t-1}| \\ &+ \mathrm{tr}\big[\{\hat{\varepsilon}_t\hat{\varepsilon}_t' + \mathrm{Var}(\varepsilon_t|Y_n)\}H_t^{-1}\big] \\ &+ \mathrm{tr}\big[\{\hat{\eta}_{t-1}\hat{\eta}_{t-1}' + \mathrm{Var}(\eta_{t-1}|Y_n)\}Q_{t-1}^{-1}\big]\big|\psi\big], \quad (7.17)\end{aligned}$$

where $\hat{\varepsilon}_t$, $\hat{\eta}_{t-1}$, $\mathrm{Var}(\varepsilon_t|Y_n)$ and $\mathrm{Var}(\eta_{t-1}|Y_n)$ are computed assuming $\psi = \tilde{\psi}$, while H_t and Q_{t-1} retain their original dependence on ψ. The equations obtained by setting (7.17) equal to zero are then solved for the elements of ψ to obtain a revised estimate of ψ. This is taken as the new trial value of ψ and the process is repeated either until adequate convergence is achieved or a switch is made to numerical maximisation of $\log L(Y_n|\psi)$. The latter option is often used since although the EM algorithm usually converges fairly rapidly in the early stages, its rate of convergence near the maximum is frequently substantially slower than numerical maximisation; see Watson and Engle (1983) and Harvey and Peters (1984) for discussion of this point. As for the score vector in the previous section, when α_1 is diffuse we merely redefine η_0 and Q_0 in such a way that they are consistent with the initial state vector model $\alpha_1 = a + A\delta + R_0\eta_0$ where $\delta \sim \mathrm{N}(0, \kappa I_q)$ and $\eta_0 \sim \mathrm{N}(0, Q_0)$ and we ignore the part associated with δ. When α_1 consists only of diffuse elements the term in Q_0^{-1} disappears from (7.17).

To illustrate, we apply the EM algorithm to the local level model as in the previous section but now we take

$$\psi = \begin{pmatrix} \sigma_\varepsilon^2 \\ \sigma_\xi^2 \end{pmatrix},$$

as the unknown parameter vector. The E-step involves the Kalman filter and disturbance smoother to obtain $\hat{\varepsilon}_t$, $\hat{\xi}_{t-1}$, $\mathrm{Var}(\varepsilon_t|Y_n)$ and $\mathrm{Var}(\xi_{t-1}|Y_n)$ of (7.17) given $\psi = \tilde{\psi}$. The M-step solves for σ_ε^2 and σ_ξ^2 by equating (7.17) to zero. For example, in a similar way as in the previous section we have

$$\begin{aligned}-2\frac{\partial}{\partial\sigma_\varepsilon^2}\tilde{\mathrm{E}}[\log p(\alpha, Y_n|\psi)] &= \frac{\partial}{\partial\sigma_\varepsilon^2}\sum_{t=1}^{n}\left[\log\sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2}\left\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\right\}\right] \\ &= \frac{n}{\sigma_\varepsilon^2} - \frac{1}{\sigma_\varepsilon^4}\sum_{t=1}^{n}\left\{\hat{\varepsilon}_t^2 + \mathrm{Var}(\varepsilon_t|Y_n)\right\} \\ &= 0,\end{aligned}$$

and similarly for the term in σ_ξ^2. New trial values for σ_ε^2 and σ_ξ^2 are therefore obtained from

$$\bar{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{t=1}^{n}\left\{\hat{\varepsilon}_t^2 - \mathrm{Var}(\varepsilon_t|Y_n)\right\} = \tilde{\sigma}_\varepsilon^2 + \frac{1}{n}\tilde{\sigma}_\varepsilon^4\sum_{t=1}^{n}\left(u_t^2 - D_t\right),$$

$$\bar{\sigma}_\xi^2 = \frac{1}{n-1}\sum_{t=2}^{n}\left\{\hat{\xi}_{t-1}^2 - \mathrm{Var}(\xi_{t-1}|Y_n)\right\} = \tilde{\sigma}_\xi^2 + \frac{1}{n-1}\tilde{\sigma}_\xi^4\sum_{t=2}^{n}\left(r_{t-1}^2 - N_{t-1}\right),$$

since $\hat{\varepsilon}_t = \tilde{\sigma}_\varepsilon^2 u_t$, $\mathrm{Var}(\varepsilon_t|Y_n) = \tilde{\sigma}_\varepsilon^2 - \tilde{\sigma}_\varepsilon^4 D_t$, $\hat{\xi}_t = \tilde{\sigma}_\xi^2 r_t$ and $\mathrm{Var}(\xi_t|Y_n) = \tilde{\sigma}_\xi^2 - \tilde{\sigma}_\xi^4 N_t$. The disturbance smoothing values u_t, D_t, r_t and N_t are based on $\tilde{\sigma}_\varepsilon^2$ and $\tilde{\sigma}_\xi^2$. The new values $\bar{\sigma}_\varepsilon^2$ and $\bar{\sigma}_\xi^2$ replace $\tilde{\sigma}_\varepsilon^2$ and $\tilde{\sigma}_\xi^2$ and the procedure is repeated until either convergence has been attained or until a switch is made to numerical optimisation. Similar elegant results are obtained for more general time series models where unknown parameters occur only in the H_t and Q_t matrices.

7.3.5 Estimation when dealing with diffuse initial conditions

It was shown in previous sections that only minor adjustments are required for parameter estimation when dealing with a diffuse initial state vector. The diffuse loglikelihood requires either the exact initial Kalman filter or the augmented Kalman filter. In both cases the diffuse loglikelihood is calculated in much the same way as for the nondiffuse case. No real new complications arise when computing the score vector or when estimating parameters via the EM algorithm. There is a compelling argument however for using the exact initial Kalman filter of Section 5.2 rather than the augmented Kalman filter of Section 5.7 for the estimation of parameters. For most practical models, the matrix $P_{\infty,t}$ and its associated matrices $F_{\infty,t}$, $M_{\infty,t}$ and $K_{\infty,t}$ do not depend on parameter vector ψ. This may be surprising but, for example, by studying the illustration given in Subsection 5.6.1 for the local linear trend model we see that the matrices $P_{\infty,t}$, $K_t^{(0)} = T_t M_{\infty,t} F_{\infty,t}^{-1}$ and $L_t^{(0)} = T_t - K_t^{(0)} Z_t$ do not depend on σ_ε^2, σ_ξ^2 or on σ_ζ^2. On the other hand, we see that all the matrices reported in Subsection 5.7.4, which deals with the augmentation approach to the same example, depend on $q_\xi = \sigma_\xi^2/\sigma_\varepsilon^2$ and $q_\zeta = \sigma_\zeta^2/\sigma_\varepsilon^2$. Therefore, every time that the parameter vector ψ changes during the estimation process we need to recalculate the augmented part of the augmented Kalman filter while we do not have to recalculate the matrices related to $P_{\infty,t}$ for the exact initial Kalman filter.

First we consider the case where only the system matrices H_t, R_t and Q_t depend on the parameter vector ψ. The matrices $F_{\infty,t} = Z_t P_{\infty,t} Z_t'$ and $M_{\infty,t} = P_{\infty,t} Z_t'$ do not depend on ψ since the update equation for $P_{\infty,t}$ is given by

$$P_{\infty,t+1} = T_t P_{\infty,t}\left(T_t - K_t^{(0)} Z_t\right)',$$

where $K_t^{(0)} = T_t M_{\infty,t} F_{\infty,t}^{-1}$ and $P_{\infty,1} = AA'$, for $t = 1, \dots, d$. Thus for all quantities related to $P_{\infty,t}$ the parameter vector ψ does not play a role. The same holds for computing a_{t+1} for $t = 1, \ldots, d$ since

$$a_{t+1} = T_t a_t + K_t^{(0)} v_t,$$

where $v_t = y_t - Z_t a_t$ and $a_1 = a$. Here again no quantity depends on ψ. The update equation

$$P_{*,t+1} = T_t P_{*,t}\big(T_t - K_t^{(0)} Z_t\big)' - K_t^{(0)} F_{\infty,t} K_t^{(1)\prime} + R_t Q_t R_t',$$

where $K_t^{(1)} = T_t M_{*,t} F_{\infty,t}^{-1} - K_t^{(0)} F_{*,t} F_{\infty,t}^{-1}$ depends on ψ. Thus we compute vector v_t and matrices $K_t^{(0)}$ and $F_{\infty,t}$ for $t = 1, \ldots, d$ once at the start of parameter estimation and we store them. When the Kalman filter is called again for likelihood evaluation we do not need to recompute these quantities and we only need to update the matrix $P_{*,t}$ for $t = 1, \ldots, d$. This implies considerable computational savings during parameter estimation using the EM algorithm or maximising the diffuse loglikelihood using a variant of Newton's method.

For the case where ψ also affects the system matrices Z_t and T_t we achieve the same computational savings for all nonstationary models we have considered in this book. The matrices Z_t and T_t may depend on ψ but the parts of Z_t and T_t which affect the computation of $P_{\infty,t}$, $F_{\infty,t}$, $M_{\infty,t}$ and $K_{\infty,t}$ for $t = 1, \ldots, d$ do not depend on ψ. It should be noted that the rows and columns of $P_{\infty,t}$ associated with elements of α_1 which are not elements of δ are zero for $t = 1, \ldots, d$. Thus the columns of Z_t and the rows and columns of T_t related to stationary elements of the state vector do not influence the matrices $P_{\infty,t}$, $F_{\infty,t}$, $M_{\infty,t}$ and $K_{\infty,t}$. In the nonstationary time series models of Chapter 3 such as the ARIMA and structural time series models, all elements of ψ which affect Z_t and T_t only relate to the stationary part of the model, for $t = 1, \ldots, d$. The parts of Z_t and T_t associated with δ only have values equal to zero and unity. For example, the ARIMA(2,1,1) model of Section 3.4 shows that $\psi = (\phi_1\,, \phi_2\,, \theta_1\,, \sigma^2)'$ does not influence the elements of Z_t and T_t associated with the first element of the state vector.

7.3.6 Large sample distribution of estimates

It can be shown that under reasonable assumptions about the stability of the model over time, the distribution of $\hat{\psi}$ for large n is approximately

$$\hat{\psi} \sim \mathrm{N}(\psi, \Omega), \tag{7.18}$$

where

$$\Omega = \left[-\frac{\partial^2 \log L}{\partial\psi\partial\psi'}\right]^{-1}. \tag{7.19}$$

This distribution has the same form as the large sample distribution of maximum likelihood estimators from samples of independent and identically distributed observations. The result (7.18) is discussed by Hamilton (1994) in Section 5.8 for general time series models and in Section 13.4 for the special case of linear Gaussian state space models. In his discussion, Hamilton gives a number of references to theoretical work on the subject.

7.3.7 Effect of errors in parameter estimation

Up to this point we have followed standard classical statistical methodology by first deriving estimates of quantities of interest on the assumption that the parameter vector ψ is known and then replacing ψ in the resulting formulae by its maximum likelihood estimate $\hat{\psi}$. We now consider the estimation of the biases in the estimates that might arise from following this procedure. Since an analytical solution in the general case seems intractable, we employ simulation. We deal with cases where $\mathrm{Var}(\hat{\psi}) = O(n^{-1})$ so the biases are also of order n^{-1}.

The technique that we propose is simple. Pretend that $\hat{\psi}$ is the true value of ψ. From (7.18) and (7.19) we know that the approximate large sample distribution of the maximum likelihood estimate of ψ given that the true ψ is $\hat{\psi}$ is $\mathrm{N}(\hat{\psi}, \hat{\Omega})$, where $\hat{\Omega}$ is Ω given by (7.19) evaluated at $\psi = \hat{\psi}$. Draw a simulation sample of N independent values $\psi^{(i)}$ from $\mathrm{N}(\hat{\psi}, \hat{\Omega})$, $i = 1, \dots, N$. Denote by e a scalar, vector or matrix quantity that we wish to estimate from the sample Y_n and let

$$\hat{e} = \mathrm{E}(e|Y_n)|_{\psi=\hat{\psi}}$$

be the estimate of e obtained by the methods of Chapter 4. For simplicity we focus on smoothed values, though an essentially identical technique holds for filtered estimates. Let

$$e^{(i)} = \mathrm{E}(e|Y_n)|_{\psi=\psi^{(i)}}$$

be the estimate of e obtained by taking $\psi = \psi^{(i)}$, for $i = 1, \dots, N$. Then estimate the bias by

$$\hat{B}_e = \frac{1}{N}\sum_{i=1}^{N} e^{(i)} - \hat{e}. \tag{7.20}$$

The accuracy of $\hat{B}_e$ can be improved significantly by the use of antithetic variables, which are discussed in detail in Subsection 11.4.3 in connection with the use of importance sampling in the treatment of non-Gaussian models. For example, we can balance the sample of $\psi^{(i)}$'s for location by taking only $N/2$ draws from $\mathrm{N}(\hat{\psi}, \hat{\Omega})$, where N is even, and defining $\psi^{(N-i+1)} = 2\hat{\psi} - \psi^{(i)}$ for $i = 1, \dots, N/2$. Since $\psi^{(N-i+1)} - \hat{\psi} = -(\psi^{(i)} - \hat{\psi})$ and the distribution of $\psi^{(i)}$ is symmetric about $\hat{\psi}$, the distribution of $\psi^{(N-i+1)}$ is the same as that of $\psi^{(i)}$. In this way we not only reduce the numbers of draws required from the $\mathrm{N}(\hat{\psi}, \hat{\Omega})$ distribution by half, but we introduce negative correlation between the $\psi^{(i)}$'s which will reduce sample variation and we have arranged the simulation sample so that the sample mean $(\psi^{(1)} + \cdots + \psi^{(N)})/N$ is equal to the population mean $\hat{\psi}$.

We can balance the sample for scale by a technique described in Subsection 11.4.3 using the fact that

$$\big(\psi^{(i)} - \hat{\psi}\big)'\hat{\Omega}^{-1}\big(\psi^{(i)} - \hat{\psi}\big) \sim \chi^2_w,$$

where w is the dimensionality of ψ; however, our expectation is that in most cases balancing for location only would be sufficient. The mean square error matrix due to simulation can be estimated in a manner similar to that described in Subsection 11.6.5.

Of course, we are not proposing that bias should be estimated as a standard part of routine time series analysis. We have included a description of this technique in order to assist workers in investigating the degree of bias in particular types of problems; in most practical cases we would expect the bias to be small enough to be neglected.

Simulation for correcting for bias due to errors in parameter estimates has previously been suggested by Hamilton (1994). His methods differ from ours in two respects. First he uses simulation to estimate the entire function under study, which in his case is a mean square error matrix, rather than just the bias, as in our treatment. Second, he has omitted a term of the same order as the bias, namely n^{-1}, as demonstrated for the local level model that we considered in Chapter 2 by Quenneville and Singh (1997). This latter paper corrects Hamilton's method and provides interesting analytical and simulation results but it only gives details for the local level model. Different methods based on parametric and nonparametric bootstrap samples have been proposed by Stoffer and Wall (1991, 2004) and Pfeffermann and Tiller (2005).

7.4 Goodness of fit

Given the estimated parameter vector $\hat{\psi}$, we may want to measure the fit of the model under consideration for the given time series. Goodness of fit measures for time series models are usually associated with forecast errors. A basic measure of fit is the forecast variance F_t which could be compared with the forecast variance of a naive model. For example, when we analyse a time series with time-varying trend and seasonal, we could compare the forecast variance of this model with the forecast variance of the time series after adjusting it with fixed trend and seasonal.

When dealing with competing models, we may want to compare the loglikelihood value of a particular fitted model, as denoted by $\log L(Y_n|\hat{\psi})$ or $\log L_d(Y_n|\hat{\psi})$, with the corresponding loglikelihood values of competing models. Generally speaking, the larger the number of parameters that a model contains the larger its loglikelihood. In order to have a fair comparison between models with different numbers of parameters, information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are used. For a univariate series they are given by

$$\text{AIC} = n^{-1}[-2\log L(Y_n|\hat{\psi}) + 2w], \qquad \text{BIC} = n^{-1}[-2\log L(Y_n|\hat{\psi}) + w\log n],$$

and with diffuse initialisation they are given by

$$\text{AIC} = n^{-1}[-2\log L_d(Y_n|\hat{\psi}) + 2(q+w)],$$
$$\text{BIC} = n^{-1}[-2\log L_d(Y_n|\hat{\psi}) + (q+w)\log n],$$

where w is the dimension of ψ. Models with more parameters or more nonstationary elements obtain a larger penalty. More details can be found in Harvey (1989). In general, a model with a smaller value of AIC or BIC is preferred.

7.5 Diagnostic checking

The diagnostic statistics and graphics discussed in Section 2.12 for the local level model (2.3) can be used in the same way for all univariate state space models. The basic diagnostics of Subsection 2.12.1 for normality, heteroscedasticity and serial correlation are applied to the one-step ahead forecast errors defined in (4.13) after standardisation by dividing by the standard deviation $F_t^{1/2}$. In the case of multivariate models, we can consider the standardised individual elements of the vector

$$v_t \sim \text{N}(0, F_t), \qquad t = d+1, \dots, n,$$

but the individual elements are correlated since matrix F_t is not diagonal. The innovations can be transformed such that they are uncorrelated:

$$v_t^s = B_t v_t, \qquad F_t^{-1} = B_t' B_t.$$

It is then appropriate to apply the basic diagnostics to the individual elements of v_t^s. Another possibility is to apply multivariate generalisations of the diagnostic tests to the full vector v_t^s. A more detailed discussion on diagnostic checking can be found in Harvey (1989) and throughout the *STAMP* manual of Koopman, Harvey, Doornik and Shephard (2010).

Auxiliary residuals for the general state space model are constructed by

$$\hat{\varepsilon}_t^s = B_t^\varepsilon \hat{\varepsilon}_t, \qquad [\text{Var}(\hat{\varepsilon}_t)]^{-1} = B_t^{\varepsilon\prime} B_t^\varepsilon,$$
$$\hat{\eta}_t^s = B_t^\eta \hat{\eta}_t, \qquad [\text{Var}(\hat{\eta}_t)]^{-1} = B_t^{\eta\prime} B_t^\eta,$$

for $t = 1, \dots, n$. The auxiliary residual $\hat{\varepsilon}_t^s$ can be used to identify outliers in the y_t series. Large absolute values in $\hat{\varepsilon}_t^s$ indicate that the behaviour of the observed value cannot be appropriately represented by the model under consideration. The usefulness of $\hat{\eta}_t^s$ depends on the interpretation of the state elements in α_t implied by the design of the system matrices T_t, R_t and Q_t. The way these auxiliary residuals can be exploited depends on their interpretation. For the local level model considered in Subsections 7.3.3 and 7.3.4 it is clear that the state is the time-varying level and ξ_t is the change of the level for time $t+1$.

It follows that structural breaks in the series y_t can be identified by detecting large absolute values in the series for $\hat{\xi}_t^s$. In the same way, for the univariate local linear trend model (3.2), the second element of $\hat{\xi}_t^s$ can be exploited to detect slope changes in the series y_t. Harvey and Koopman (1992) have formalised these ideas further for the structural time series models of Section 3.2 and they constructed some diagnostic normality tests for the auxiliary residuals.

It is argued by de Jong and Penzer (1998) that such auxiliary residuals can be computed for any element of the state vector and that they can be considered as t-tests for the hypotheses

$$H_0 : (\alpha_{t+1} - T_t\alpha_t - R_t\eta_t)_i = 0,$$

the appropriate large-sample statistic for which is computed by

$$r_{it}^s = r_{it}/\sqrt{N_{ii,t}},$$

for $i = 1, \ldots, m$, where $(\cdot)_i$ is the ith element of the vector within brackets, r_{it} is the ith element of the vector r_t and $N_{ij,t}$ is the (i, j)th element of the matrix N_t; the recursions for evaluating r_t and N_t are given in Subsection 4.5.3. The same applies to the measurement equation for which t-test statistics for the hypotheses

$$H_0 : (y_t - Z_t\alpha_t - \varepsilon_t)_i = 0,$$

are computed by

$$e_{it}^s = e_{it}/\sqrt{D_{ii,t}},$$

for $i = 1, \ldots, p$, where the equations for computing e_t and D_t are given in Subsection 4.5.3. These diagnostics can be regarded as model specification tests. Large values in r_{it}^s and e_{it}^s, for some values of i and t, may reflect departures from the overall model and they may indicate specific adjustments to the model.

8 Illustrations of the use of the linear model

8.1 Introduction

In this chapter we will give some illustrations which show how the use of the linear model works in practice. State space methods are usually employed for time series problems and most of our examples will be from this area but we also will treat a smoothing problem which is not normally regarded as part of time series analysis and which we solve using cubic splines.

The first example is an analysis of road accident data to estimate the reduction in car drivers killed and seriously injured in the UK due to the introduction of a law requiring the wearing of seat belts. In the second example we consider a bivariate model in which we include data on numbers of front seat passengers killed and seriously injured and on numbers of rear seat passengers killed and seriously injured and we estimate the effect that the inclusion of the second variable has on the accuracy of the estimation of the drop in the first variable. The third example shows how state space methods can be applied to Box–Jenkins ARMA models employed to model series of users logged onto the Internet. In the fourth example we consider the state space solution to the spline smoothing of motorcycle acceleration data. The fifth example provides a dynamic factor analysis based on the linear Gaussian model for the term structure of interest rates paid on US Treasury securities. The software we have used for most of the calculations is *SsfPack* and is described in Section 6.7.

8.2 Structural time series models

The study by Durbin and Harvey (1985) and Harvey and Durbin (1986) on the effect of the seat belt law on road accidents in Great Britain provides an illustration of the use of structural time series models for the treatment of problems in applied time series analysis. They analysed data sets which contained numbers of casualties in various categories of road accidents to provide an independent assessment on behalf of the Department of Transport of the effects of the British seat belt law on road casualties. Most series were analysed by means of linear Gaussian state space models. We concentrate here on monthly numbers of drivers, front seat passengers and rear seat passengers who were killed or seriously injured in road accidents in cars in Great Britain from January 1969 to December 1984. Data were transformed into logarithms since logged values fitted the model better. Data on the average number of kilometres travelled per car per month and the real price of petrol are included as possible explanatory variables. We start with a univariate analysis of the drivers series. In the next section we perform a bivariate analysis using the front and rear seat passengers.

The log of monthly number of car drivers killed or seriously injured is displayed in Fig. 8.1. The graph shows a seasonal pattern which may be due to weather conditions and festive celebrations. The overall trend of the series is basically constant over the years with breaks in the mid-1970s, probably due to the oil crisis, and in February 1983 after the introduction of the seat belt law. The model that we shall consider initially is the basic structural time series model which is given by

$$y_t = \mu_t + \gamma_t + \varepsilon_t,$$

where μ_t is the local level component modelled as the random walk $\mu_{t+1} = \mu_t + \xi_t$, γ_t is the trigonometric seasonal component (3.7) and (3.8), and ε_t is a disturbance term with mean zero and variance σ^2_ε. Note that for illustrative purposes we do not at this stage include an intervention component to measure the effect of the seat belt law.

The model is estimated by maximum likelihood using the techniques described in Chapter 7. The iterative method of finding the estimates for σ^2_ε,

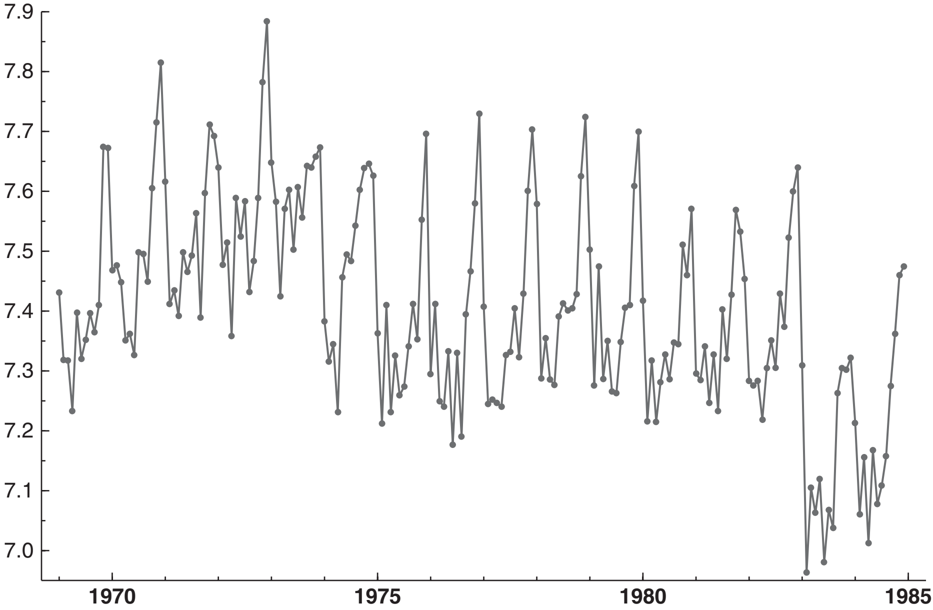

Fig. 8.1 Monthly numbers (logged) of drivers who were killed or seriously injured (KSI) in road accidents in cars in Great Britain.

σ^2_ξ and σ^2_ω is implemented in *STAMP* 8.3 based on the concentrated diffuse loglikelihood as discussed in Subsection 2.10.2; the estimation output is given below where the first element of the parameter vector is $\phi_1 = 0.5 \log q_\eta$ and the second element is $\phi_2 = 0.5 \log q_\omega$ where $q_\xi = \sigma^2_\xi / \sigma^2_\varepsilon$ and $q_\omega = \sigma^2_\omega / \sigma^2_\varepsilon$. We present the parameter estimation results obtained from the *STAMP* package of Koopman, Harvey, Doornik and Shephard (2010). The estimates are obtained from a small number of cycles of univariate optimisations with respect to one parameter and for which the other parameters are kept fixed to their current values, starting with arbitrary values for the univariate optimisations. The initial parameter estimates obtained from this procedure are usually good starting values for the simultaneous optimisation procedure that produces the maximum likelihood estimates. In this procedure, one variance parameter is concentrated out. The resulting parameter estimates are given below. The estimate for σ^2_ω is very small but when it is set equal to zero, evidence of seasonal serial correlation is found in the residuals. Therefore we keep σ^2_ω equal to its estimated value.

```
Estimated variances of disturbances

Component                    Value    (q-ratio)
Level                  0.000935852    ( 0.2740)
Seasonal              5.01096e-007  (0.0001467)
Irregular               0.00341598     ( 1.000)
```

The estimated components are displayed in Fig. 8.2. The estimated level does pick up the underlying movements in the series and the estimated irregular does not cause much concern to us. The seasonal effect hardly changes over time.

In Fig. 8.3 the estimated level is displayed; the predicted estimator is based on only the past data, that is $\mathrm{E}(\mu_t | Y_{t-1})$, and the smoothed estimator is based on all the data, that is $\mathrm{E}(\mu_t | Y_n)$. It can be seen that the predicted estimator lags the shocks in the series as is to be expected since this estimator does not take account of current and future observations.

The model fit of this series and the basic diagnostics initially appear satisfactory. The standard output provided by *STAMP* is given by

```
Diagnostic summary report

Estimation sample is 69. 1 - 84.12. (T = 192, n = 180).
Log-Likelihood is 435.295 (-2 LogL = -870.59).
Prediction error variance is 0.00586717
Summary statistics
             drivers
Std.Error  0.076597
N            4.6692
H( 60)       1.0600
r( 1)      0.038621
Q(24,22)     33.184
```

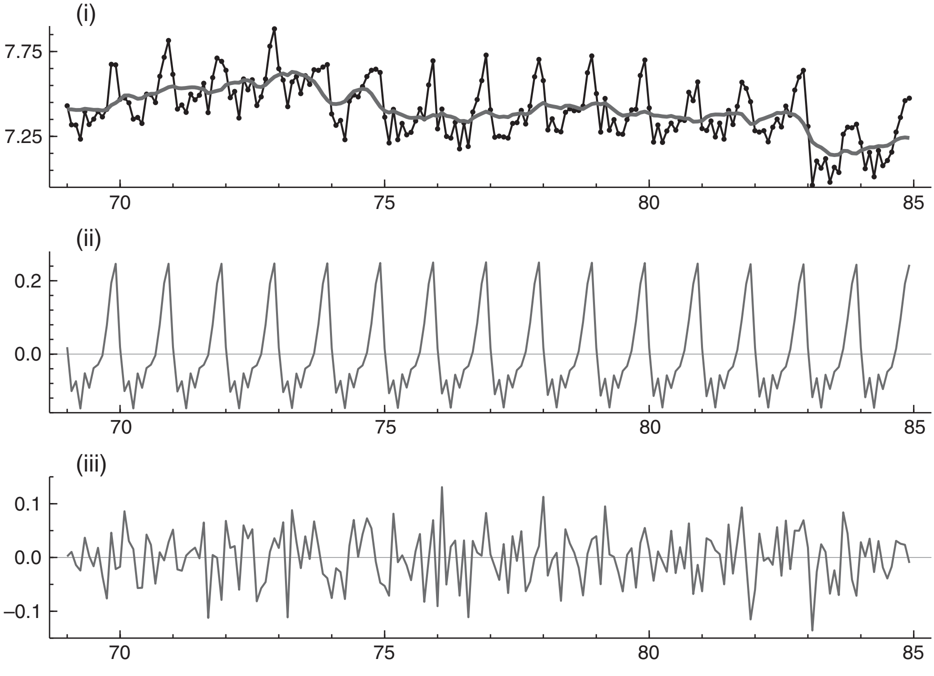

Fig. 8.2 Estimated components: (i) level; (ii) seasonal; (iii) irregular.

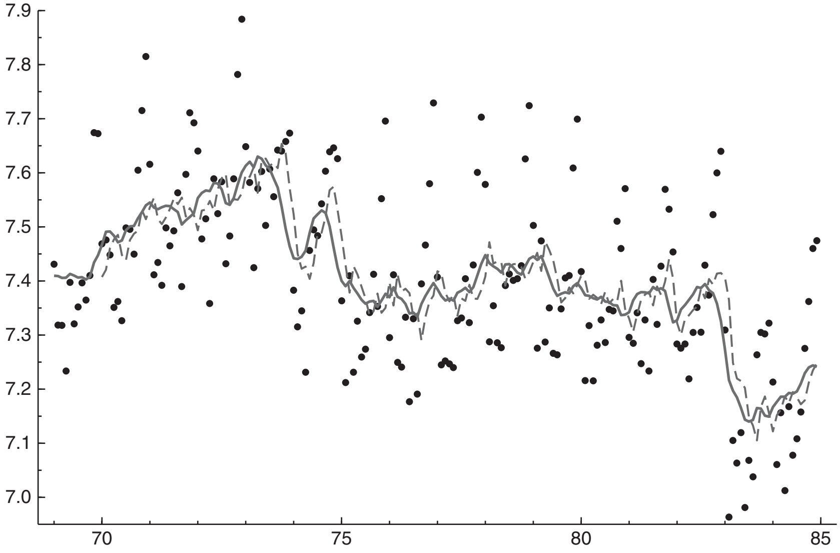

Fig. 8.3 Data (dots) with predicted (dashed line) and smoothed (solid line) estimated level.

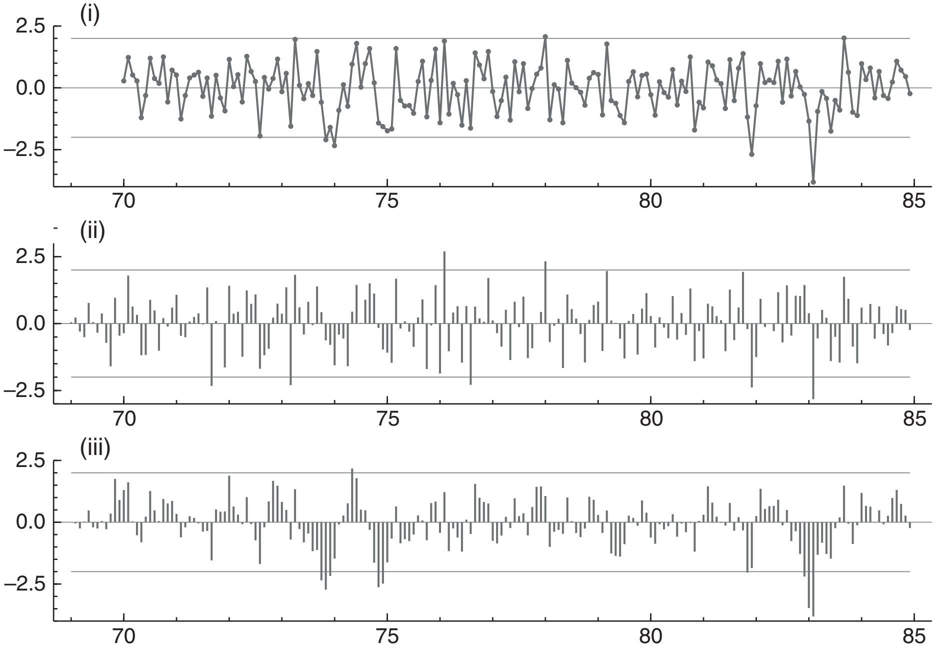

Fig. 8.4 (i) The one-step ahead prediction residuals (time series plot); (ii) auxiliary irregular residuals (index plot); (iii) auxiliary level residuals (index plot).

The definitions of the diagnostics can be found in Section 2.12.

When we inspect the graphs of the residuals in Fig. 8.4, however, in particular the auxiliary level residuals, we see a large negative value for February 1983. This suggests a need to incorporate an intervention variable to measure the level shift in February 1983. We have performed this analysis without inclusion of such a variable purely for illustrative purposes; obviously, in a real analysis the variable would be included since a drop in casualties was expected to result from the introduction of the seat belt law.

By introducing an intervention which equals one from February 1983 and is zero prior to that and the price of petrol as a further explanatory variable, we re-estimate the model and obtain the regression output

```
Estimated coefficients of explanatory variables.

Variable   Coefficient  R.m.s.e.  t-value
petrol     -0.29140     0.09832   -2.96384 [ 0.00345]
Lvl 83. 2  -0.23773     0.04632   -5.13277 [ 0.00000]
```

The estimated components, when the intervention variable and the regression effect due to the price of petrol are included, are displayed in Fig. 8.5.

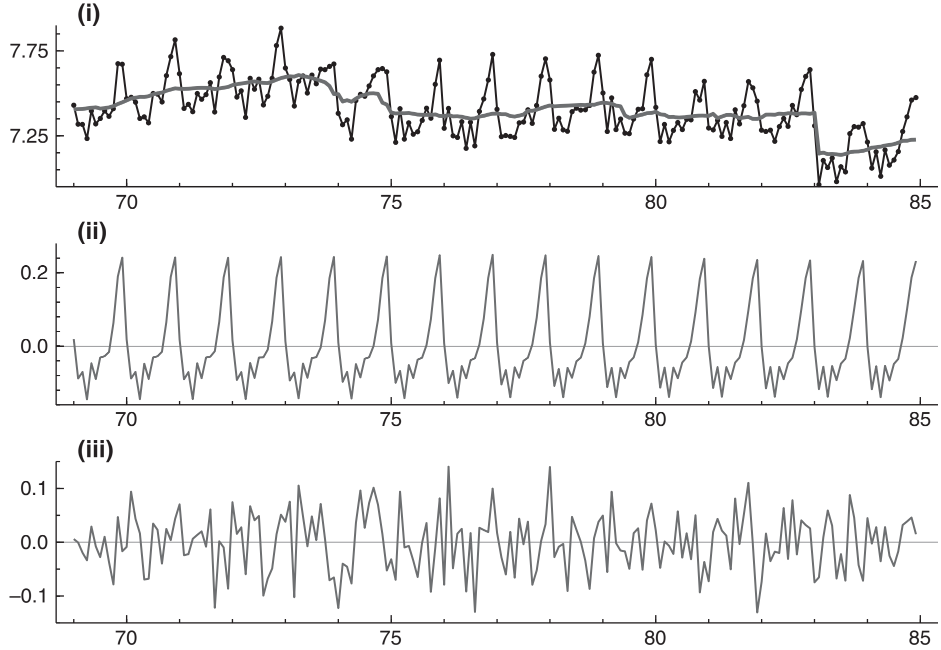

Fig. 8.5 Estimated components for model with intervention and regression effects: (i) level; (ii) seasonal; (iii) irregular.

The estimated coefficient of a break in the level after January 1983 is -0.238, indicating a fall of 21%, that is, $1 - \exp(-0.238) = 0.21$, in car drivers killed and seriously injured after the introduction of the law. The t-value of -5.1 indicates that the break is highly significant. The coefficient of petrol price is also significant.

8.3 Bivariate structural time series analysis

Multivariate structural time series models are introduced in Section 3.3. To illustrate state space methods for a multivariate model we analyse a bivariate monthly time series of front seat passengers and rear seat passengers killed and seriously injured in road accidents in Great Britain which were included in the assessment study by Harvey and Durbin (1986).

The graphs in Fig. 8.6 indicate that the local level specification is appropriate for the trend component and that we need to include a seasonal component. We start by estimating a bivariate model with level, trigonometric seasonal and irregular components together with explanatory variables for the number of kilometres travelled and the real price of petrol. We estimate the model only using

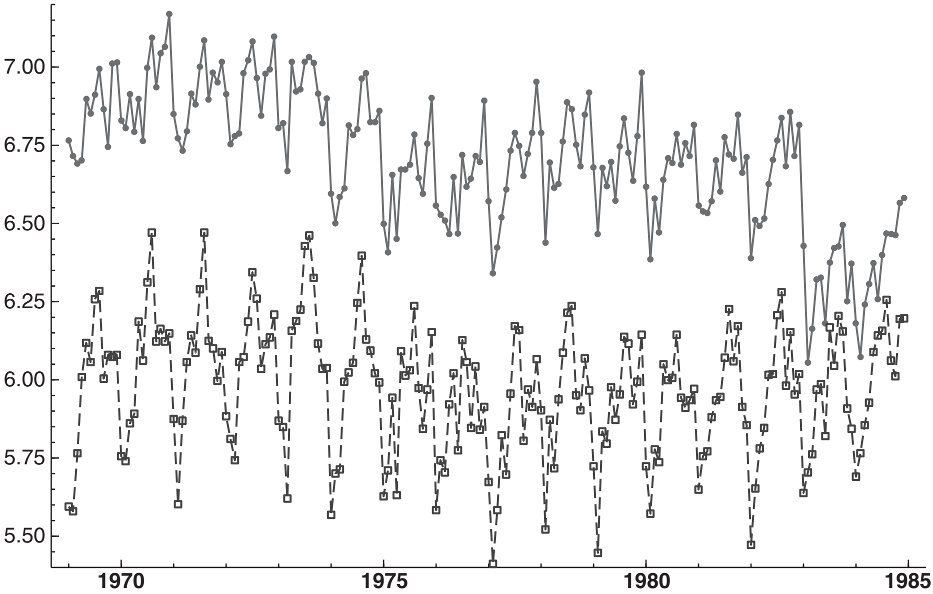

Fig. 8.6 Front seat (grey, dotted line) and rear seat (dashed line with squares) passengers killed and seriously injured in road accidents in Great Britain.

observations available before 1983, the year in which the seat belt law was introduced. The variance matrices of the three disturbance vectors are estimated by maximum likelihood in the way described in Section 7.3. The estimated variance matrix of the seasonal component is small which lead us to model the seasonal component as fixed and to re-estimate the remaining two variances matrices:

$$\widehat{\Sigma}_\varepsilon = 10^{-4}\begin{bmatrix}5.006 & 4.569\\ 4.569 & 9.143\end{bmatrix},\ \widehat{\Sigma}_\eta = 10^{-5}\begin{bmatrix}4.834 & 2.993\\ 2.993 & 2.234\end{bmatrix},\ \hat{\rho}_\varepsilon = 0.675,\ \hat{\rho}_\eta = 0.893,$$

where $\rho_x = \Sigma_x(1,2)\,/\,\sqrt{\Sigma_x(1,1)\,\Sigma_x(2,2)}$ for $x = \varepsilon, \eta$ and where $\Sigma_x(i,j)$ is the (i,j) element of matrix Σ_x for $i,j = 1,2$. The loglikelihood value of the estimated model is 742.088 with AIC equal to -4.4782.

The correlation between the two level disturbances is close to one. It may therefore be interesting to re-estimate the model with the restriction that the rank of the level variance matrix is one:

$$\widehat{\Sigma}_\varepsilon = 10^{-4}\begin{bmatrix}5.062 & 4.791\\ 4.791 & 10.02\end{bmatrix},\ \widehat{\Sigma}_\eta = 10^{-5}\begin{bmatrix}4.802 & 2.792\\ 2.792 & 1.623\end{bmatrix},\ \hat{\rho}_\varepsilon = 0.673,\ \rho_\eta = 1,$$

The loglikelihood value of this estimated model is 739.399 with AIC equal to -4.4628. A comparison of the two AIC's shows only a marginal preference for the unrestricted model.

We now assess the effect of the introduction of the seat belt law as we have done for the drivers series using a univariate model in Section 8.2. We concentrate on the effect of the law on front seat passengers. We also have the rear seat series which is highly correlated with the front seat series. However, the law did not apply to rear seat passengers so the data should therefore not be affected by the introduction of the law. Under such circumstances the rear seat series may be used as a *control group* which may result in a more precise measure of the effect of the seat belt law on front seat passengers; for the reasoning behind this idea see the discussion by Harvey (1996) to whom this approach is owed.

We consider the unrestricted bivariate model but with a level intervention for February 1983 added to both series. This model is estimated using the whole data set giving the parameter estimates:

$$\widehat{\Sigma}_\varepsilon = 10^{-4}\begin{bmatrix} 5.135 & 4.493 \\ 4.493 & 9.419 \end{bmatrix}, \ \widehat{\Sigma}_\eta = 10^{-5}\begin{bmatrix} 4.896 & 3.025 \\ 3.025 & 2.317 \end{bmatrix}, \ \hat{\rho}_\varepsilon = 0.660, \ \hat{\rho}_\eta = 0.898,$$

The estimates for the level intervention coefficient in both equation are given by

	coeff	rmse	t-value	p-value
front seat	−0.32799	0.05699	−5.75497	0.0000
rear seat	0.03376	0.05025	0.67189	0.50252

From these results and the time series plot of casualties in rear seat passengers in Fig. 8.6, it is clear that they are unaffected by the introduction of the seat belt as we expect. By removing the intervention effect from the rear seat equation of the model we obtain the estimation results:

$$\widehat{\Sigma}_\varepsilon = 10^{-4}\begin{bmatrix} 5.147 & 4.588 \\ 4.588 & 9.380 \end{bmatrix}, \ \widehat{\Sigma}_\eta = 10^{-5}\begin{bmatrix} 4.754 & 2.926 \\ 2.926 & 2.282 \end{bmatrix}, \ \hat{\rho}_\varepsilon = 0.660, \ \hat{\rho}_\eta = 0.888,$$

with the level intervention estimate given by

	coeff	rmse	t-value	p-value
front seat	−0.35630	0.03655	−9.74877	0.0000

The almost two-fold decrease of the root mean squared error for the estimated intervention coefficient for the front seat series is remarkable. Enforcing the rank of Σ_η to be one, such that the levels are proportional to each other, produces the following estimation results:

$$\widehat{\Sigma}_\varepsilon = 10^{-4}\begin{bmatrix} 5.206 & 4.789 \\ 4.789 & 10.24 \end{bmatrix}, \ \widehat{\Sigma}_\eta = 10^{-5}\begin{bmatrix} 4.970 & 2.860 \\ 2.860 & 1.646 \end{bmatrix}, \ \hat{\rho}_\varepsilon = 0.659, \ \rho_\eta = 1,$$

with level intervention estimate given by

	coeff	rmse	t-value	p-value
front seat	−0.41557	0.02621	−15.85330	0.0000

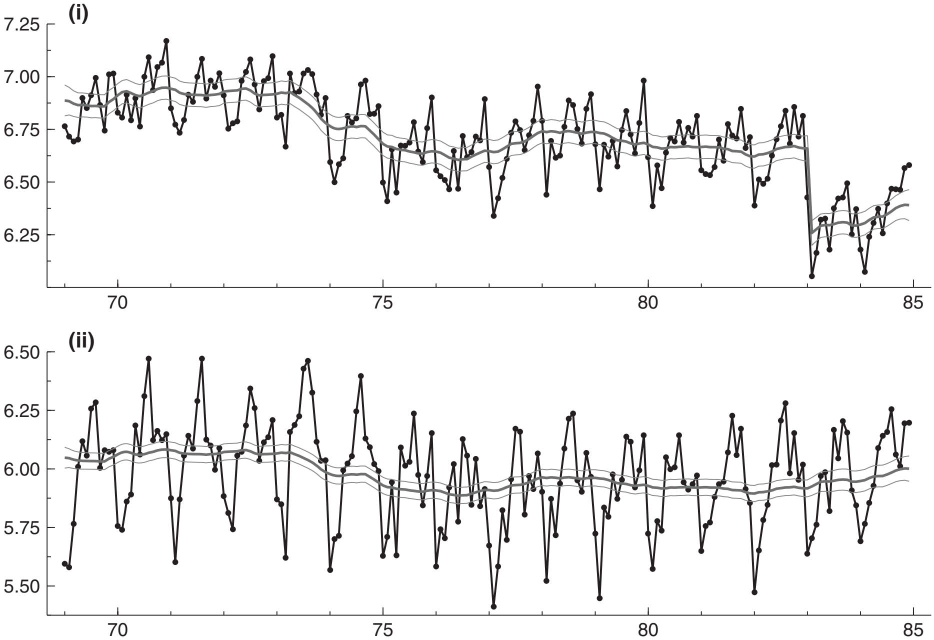

Fig. 8.7 (i) Front seat passengers with estimated signal (without seasonal) and 95% confidence interval; (ii) Rear seat passengers with estimated signal (without seasonal) and 95% confidence interval.

The rank reduction also leads to a large increase (in absolute value) of the t-value. The graphs of the estimated (non-seasonal) signals and the estimated levels for the last model are presented in Fig. 8.7. The substantial drop of the underlying level in front seat passenger casualties at the introduction of the seat belt law is clearly visible.

8.4 Box–Jenkins analysis

In this section we will show that fitting of ARMA models, which is an important part of the Box–Jenkins methodology, can be done using state space methods. Moreover, we will show that missing observations can be handled within the state space framework without problems whereas this is difficult within the Box–Jenkins methodology; see the discussion in Subsection 3.10.1. Finally, since an important objective of the Box–Jenkins methodology is forecasting, we also present forecasts of the series under investigation. In this illustration we use the series which is analysed by Makridakis, Wheelwright and Hyndman (1998): the number of users logged on to an Internet server each minute over 100 minutes.

The data are differenced in order to get them closer to stationarity and these 99 observations are presented in Fig. 8.8(i).

We have estimated a range of ARMA model (3.17) with different choices for p and q. They were estimated in state space form based on (3.20). Table 8.1 presents the Akaike information criteria (AIC), which is defined in Section 7.4, for these different ARMA models. We see that the ARMA models with (p, q) equal to $(1, 1)$ and $(3, 0)$ are optimal according to the AIC values. We prefer the ARMA$(1, 1)$ model because it is more parsimonious. A similar table was produced for the same series by Makridakis, Wheelwright and Hyndman (1998) but the AIC statistic was computed differently. They concluded that the ARMA$(3, 0)$ model was best.

We repeat the calculations for the same differenced series but now with 14 observations treated as missing: $6, 16, 26, 36, 46, 56, 66, 72, 73, 74, 75, 76, 86, 96$. The graph of the amended series is produced in Fig. 8.8(ii). The reported AIC's in Table 8.2 lead to the same conclusion as for the series without missing observations: the preferred model is ARMA$(1, 1)$ although its case is less strong now. We also learn from this illustration that estimation of higher order ARMA models with missing observations lead to more numerical problems.

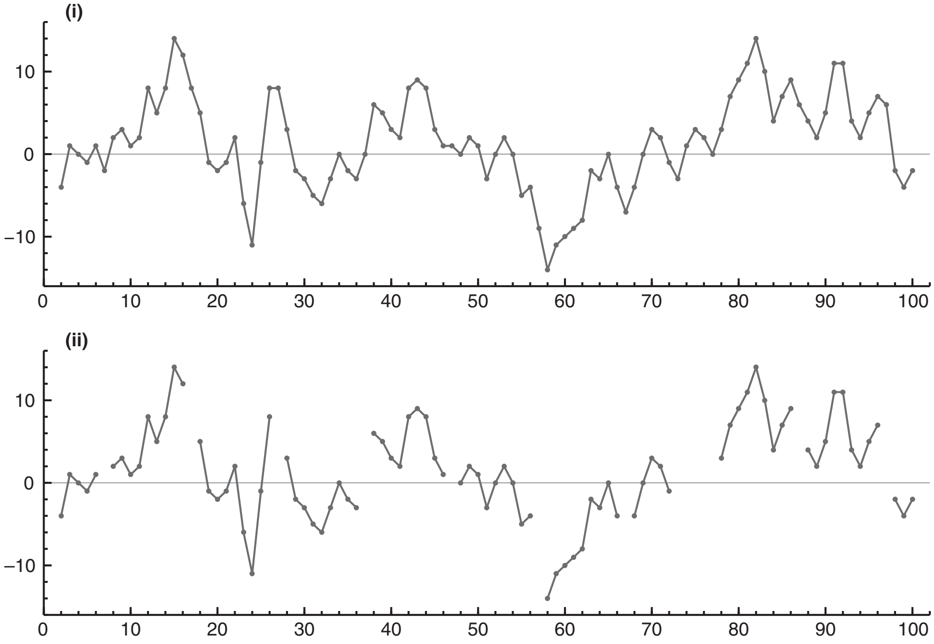

Fig. 8.8 (i) First difference of number of users logged on to Internet server each minute; (ii) The same series with 14 observations omitted.

Table 8.1 AIC for different ARMA models.

q	0	1	2	3	4	5
p						
0		2.777	2.636	2.648	2.653	2.661
1	2.673	2.608	2.628	2.629 (1)	2.642	2.658
2	2.647	2.628	2.642	2.657	2.642 (1)	2.660 (4)
3	2.606	2.626	2.645	2.662	2.660 (2)	2.681 (4)
4	2.626	2.646 (8)	2.657	2.682	2.670 (1)	2.695 (1)
5	2.645	2.665 (2)	2.654 (9)	2.673 (10)	2.662 (12)	2.727 (A)

The value between parentheses indicates the number of times the loglikelihood could not be evaluated during optimisation. The symbol A indicates that the maximisation process was automatically aborted due to numerical problems.

Table 8.2 AIC for different ARMA models with missing observations.

q	0	1	2	3	4	5
p						
0		3.027	2.893	2.904	2.908	2.926
1	2.891	2.855	2.877	2.892	2.899	2.922
2	2.883	2.878	2.895 (6)	2.915	2.912	2.931
3	2.856 (1)	2.880	2.909	2.924	2.918 (12)	2.940 (1)
4	2.880	2.901	2.923	2.946	2.943	2.957 (2)
5	2.901	2.923	2.877 (A)	2.897 (A)	2.956 (26)	2.979

The value between parentheses indicates the number of times the loglikelihood could not be evaluated during optimisation. The symbol A indicates that the maximisation process was automatically aborted due to numerical problems.

Finally, we present in Fig. 8.9 the in-sample one-step ahead forecasts for the series with missing observations and the out-of-sample forecasts with 50% confidence intervals. It is one of the many advantages of state space modelling that it allows for missing observations without difficulty.

8.5 Spline smoothing

The connection between smoothing splines and the local linear trend model has been known for many years; see, for example, Wecker and Ansley (1983). In Section 3.9 we showed that the equivalence is with a local linear trend formulated in continuous time with the variance of the level disturbance equal to zero.

Consider a set of observations $y_1, \ldots, y_n$ which are irregularly spaced and associated with an ordered series $\tau_1, \ldots, \tau_n$. The variable τ_t can also be a measure for age, length or income, for example, as well as time. The discrete time model implied by the underlying continuous time model is the local linear trend model with

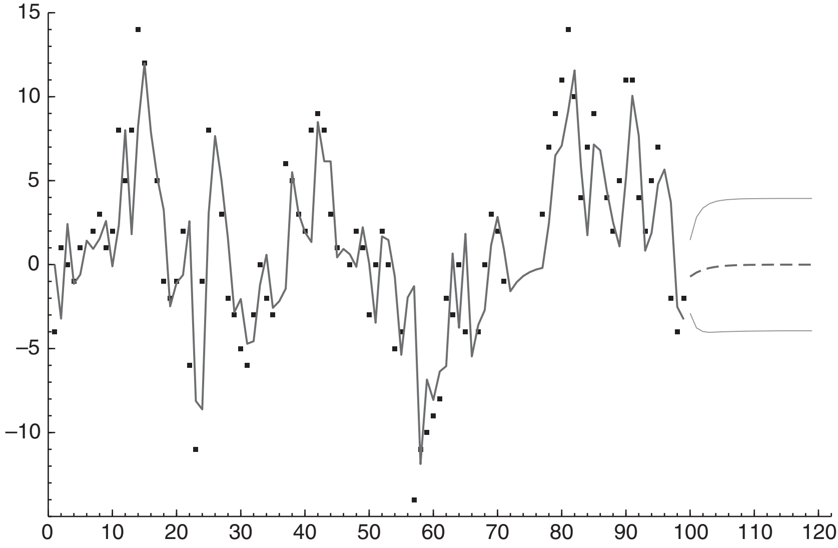

Fig. 8.9 Internet series (solid blocks) with in-sample one-step ahead predictions and out-of-sample forecasts with 50% confidence interval.

$$\mathrm{Var}(\eta_t) = \sigma_\zeta^2 \delta_t^3/3, \qquad \mathrm{Var}(\zeta_t) = \sigma_\zeta^2 \delta_t, \qquad E(\eta_t \zeta_t) = \sigma_\zeta^2 \delta_t^2/2, \qquad (8.1)$$

as shown in Section 3.8, where the distance variable $\delta_t = \tau_{t+1} - \tau_t$ is the time between observation t and observation $t+1$. We shall show how irregularly spaced data can be analysed using state space methods. With evenly spaced observations the δ_t's are set to one.

We consider 133 observations of acceleration against time (measured in milliseconds) for a simulated motorcycle accident. This data set was originally analysed by Silverman (1985) and is often used as an example of curve fitting techniques; see, for example, Hardle (1990) and Harvey and Koopman (2000). The observations are not equally spaced and at some time points there are multiple observations; see Fig. 8.10. Cubic spline and kernel smoothing techniques depend on a choice of a smoothness parameter. This is usually determined by a technique called *cross-validation.* However, setting up a cubic spline as a state space model enables the smoothness parameter to be estimated by maximum likelihood and the spline to be computed by the Kalman filter and smoother. The model can easily be extended to include other unobserved components and explanatory variables, and it can be compared with alternative models using standard statistical criteria.

We follow here the analysis given by Harvey and Koopman (2000). The smoothing parameter $\lambda = \sigma_\zeta^2/\sigma_\varepsilon^2$ is estimated by maximum likelihood (assuming normally distributed disturbances) using the transformation $\lambda = \exp(\psi)$. The

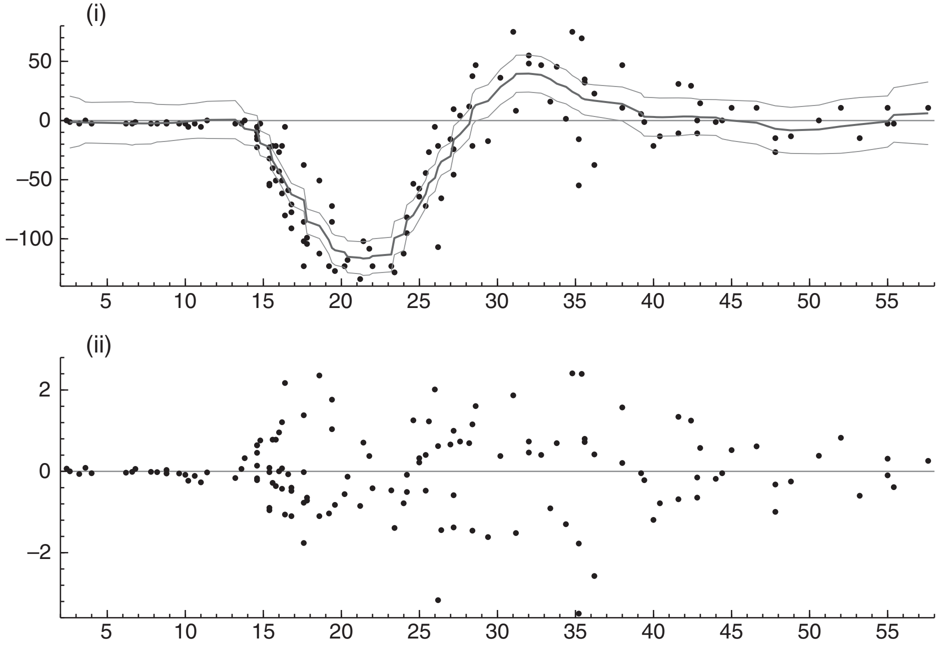

Fig. 8.10 Motorcycle acceleration data analysed by a cubic spline. (i) observations against time with spline and 95% confidence intervals, (ii) standardised irregular.

estimate of ψ is -3.59 with asymptotic standard error 0.22. This implies that the estimate of λ is 0.0275 with an asymmetric 95% confidence interval of 0.018 to 0.043. Silverman (1985) estimates λ by cross-validation, but does not report its value. In any case, it is not clear how one would compute a standard error for an estimate obtained by cross-validation. The Akaike information criterion (AIC) is 9.43. Fig. 8.10 (i) presents the cubic spline. One of the advantages of representing the cubic spline by means of a statistical model is that, with little additional computing, we can obtain variances of our estimates and, therefore, standardised residuals defined as the residuals divided by the overall standard deviation. The 95% confidence intervals for the fitted spline are also given in Fig. 8.10 (i). These are based on the root mean square errors of the smoothed estimates of μ_t, obtained from V_t as computed by (4.43), but without an allowance for the uncertainty arising from the estimation of λ as discussed in Subsection 7.3.7.

8.6 Dynamic factor analysis

The yield curve, or the term structure of interest rates, describes the relation between the interest rate (cost of borrowing) and the time to maturity of the debt for a given borrower in a given currency. The US dollar interest rates paid on US Treasury securities for various maturities are closely watched by economists and traders. The shape of the yield curve is particularly scrutinised because it provides an indication of future interest rate change and economic activity. The typical yield curve is one in which longer maturity loans have a higher yield compared to shorter-term loans due to the risks associated with time. A well-known monthly data set of yields consists of US interest rates for 17 differents maturities over the period from January 1985 up to December 2000. The maturities are 3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, 60, 72, 84, 96, 108 and 120 months. The data set is presented in Fig. 8.11. We refer to Diebold and Li (2006) for more details on this data set.

The yield curve is typically a smooth function of maturity. Nelson and Siegel (1987) propose to describe the yield curve by three factors that represent level, slope and curvature of the yield curve. Let y_{it} be the interest rate at time t for maturity τ_i for $t = 1, \dots, n$ and $i = 1, \dots, N$. The Nelson–Siegel regression model for the yield curve at time t can be expressed by

$$y_{it} = \beta_{1t} + x_{i2}\beta_{2t} + x_{i3}\beta_{3t} + \varepsilon_{it}, \tag{8.2}$$

with

$$x_{i2} = \frac{1 - z_i}{\lambda\tau_i}, \qquad x_{i3} = \frac{1 - z_i}{\lambda\tau_i} - z_i, \qquad z_i = \exp(\, -\lambda\tau_i \,),$$

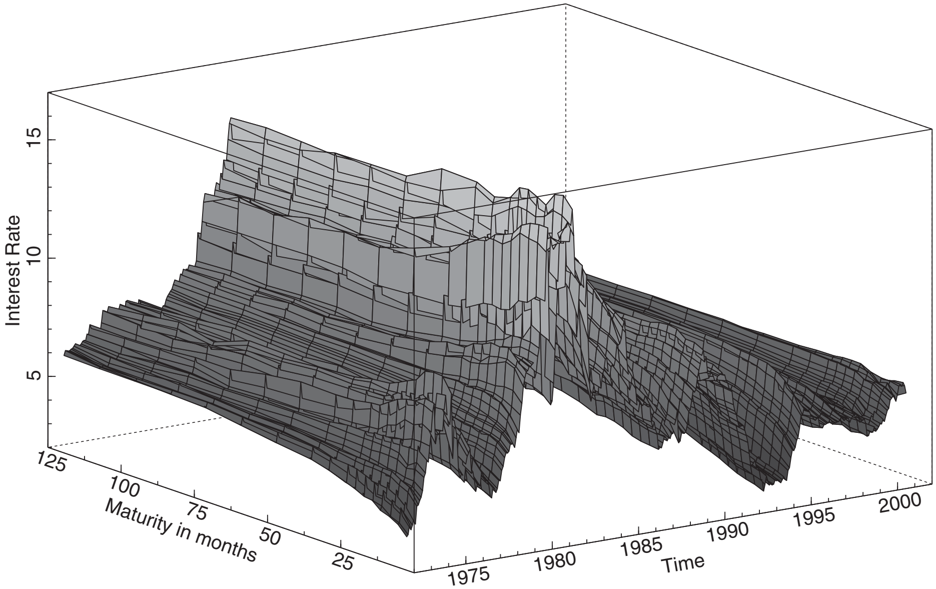

Fig. 8.11 Monthly US interest rates for maturities of 3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, 60, 72, 84, 96, 108 and 120 months, for the period from January 1985 up to December 2000.

where β_{jt} can be treated as regression coefficients, $j = 1, 2, 3$, λ is a nonlinear coefficient and ε_{it} is a normally independently distributed disturbance with mean zero and unknown variance σ^2 for $i = 1, \ldots, N$. The construction of the regressors x_{ij} allow β_{jt} to have an interpretation for $j = 1, 2, 3$ and $i = 1, \ldots, N$. The first coefficient f_{1t} represents the level or mean for all interest rates. Since regressor x_{i2} converges to one as τ_i gets smaller and converges to zero as τ_i gets larger, coefficient β_{2t} identifies the slope of the yield curve. The regressor x_{i3} converges to zero as τ_i gets smaller or larger, is concave in τ_i, and therefore coefficient β_{3t} reflects the shape of the yield curve. At time t and for yield observations $y_{1t}, \ldots, y_{Nt}$, with $N > 5$, we can estimate the unknown coefficients via nonlinear least squares. When coefficient λ is set equal to some fixed value, the remaining coefficients can be estimated by ordinary least squares.

Diebold, Rudebusch and Aruoba (2006) adopt the Nelson–Siegel framework but specify a dynamic process for the three factors and carry out a state space analysis for the resulting dynamic factor model (3.32) as given by

$$y_t = \Lambda f_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma^2 I_N),$$

where $y_t = (y_{1t}, \ldots, y_{Nt})'$, $f_t = (\beta_{1t}, \beta_{2t}, \beta_{3t})'$, $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Nt})'$ and the (ij) element of loading matrix Λ equals x_{ij}, with $x_{i1} = 1$, for $i = 1, \ldots, N$, $j = 1, 2, 3$ and $t = 1, \ldots, n$. The dynamic specification for f_t is generally given by $f_t = U_t \alpha_t$ in (3.32) where U_t is typically a known selection matrix. It allows the vector autoregressive process for the 3×1 vector f_t as proposed by Diebold, Rudebusch and Aruoba (2006) and others, that is

$$f_{t+1} = \Phi f_t + \eta_t, \qquad \eta_t \sim \mathrm{N}(0, \Sigma_\eta),$$

with $U_t = I_3$ and where Φ is the vector autoregressive coeffient matrix and Σ_η is the disturbance variance matrix. It is often assumed that vector f_t follows a stationary process. The disturbance vectors ε_t and η_t are mutually and serially uncorrelated at all times and lags. It follows that the dynamic Nelson–Siegel model is the state space model (3.1) with $Z_t = \Lambda$, $H_t = \sigma^2 I_N$, $T_t = \Phi$, $R_t = I_3$ and $Q_t = \Sigma_\eta$ for $t = 1, \ldots, n$. This dynamic factor model describes a linear relationship between the interest rates and the dynamic level, slope and curvature factors as advocated by Litterman and Scheinkman (1991).

A state space analysis includes the estimation of λ, σ^2, Φ and Σ_η by the method of maximum likelihood for which the Kalman filter is used for likelihood evaluation. Before the Kalman filter is applied, the $N \times 1$ observation vector y_t is collapsed to a 3×1 vector y_t^* as described in Section 6.5. It leads to a computationally efficient method for parameter estimation and analysing the yield curve based on the dynamic Nelson–Siegel model. The maximum likelihood estimation results are

$$\begin{array}{l}\hat{\lambda} = 0.078, \\ \hat{\sigma}^2 = 0.014,\end{array} \quad \hat{\Phi} = \begin{bmatrix} 0.994 & 0.029 & -0.022 \\ -0.029 & 0.939 & 0.040 \\ 0.025 & 0.023 & 0.841 \end{bmatrix}, \ \hat{\Sigma}_\eta = \begin{bmatrix} 0.095 & -0.014 & 0.044 \\ -0.014 & 0.383 & 0.009 \\ 0.044 & 0.009 & 0.799 \end{bmatrix}.$$

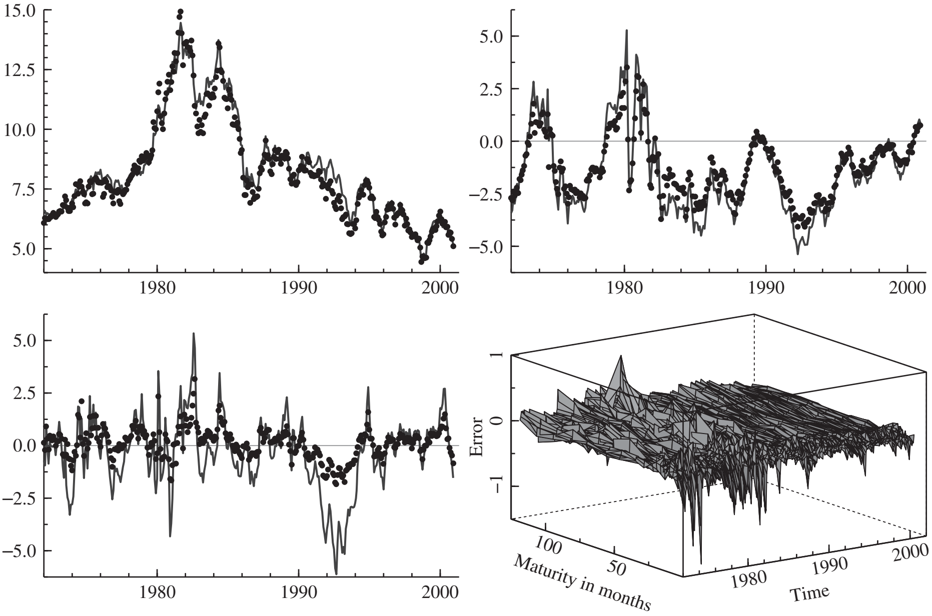

Fig. 8.12 Smoothed estimates of factors f_t and disturbances ε_t for the period from January 1985 up to December 2000: (i) data-proxy of level (dots) and $\hat{f}_{1t}$ (solid line); (ii) data-proxy of slope and $\hat{f}_{2t}$; (iii) data-proxy of slope and $\hat{f}_{3t}$; (iv) smoothed disturbances $\hat{\varepsilon}_{it}$, $i = 1, \ldots, N$ with $N = 17$.

The smoothed estimates of the time-varying factors f_t are displayed in Fig. 8.12. The three smoothed factors are displayed together with their data-proxies. An approximation of the level is the interest rate for a maturity of 10 years. The slope factor can be approximated by the interest rate spread that is defined as the difference between the rates for maturities 3 months and 10 years. The curvature proxy is the difference between the spread associated with 3 months and 2 years, and the spread associated with 10 years and 2 years. The estimated factors follow the same patterns over time in comparison to their data approximations and therefore the interpretations of the factors as level, slope and curvature of the yield curve are justified. The smoothed disturbance vectors $\hat{\varepsilon}_t$ are also displayed in a similar way as the data is presented in Fig. 8.11. We can learn that the model is more successful in fitting the long-term interest rates although most of the larger errors (in absolute values) occur in the months between the early 1970s and the early 1980s, a period of economic turmoil. An analysis with a more comprehensive model for this panel of time series is conducted by Koopman, Mallee and van der Wel (2010).

This page intentionally left blank

Part II

Non-Gaussian and nonlinear state space models

In Part I we presented a comprehensive treatment of the construction and analysis of linear Gaussian state space models, linear models and Bayesian versions of these, and we discussed the software required for implementing the related methodology. Methods based on these models, possibly after transformation of the observations, are appropriate for a wide range of problems in practical time series analysis.

There are situations, however, where the linear Gaussian model fails to provide an acceptable representation of the behaviour of the data. For example, if the observations are monthly numbers of people killed in road accidents in a particular region, and if the numbers concerned are relatively small, the Poisson distribution will generally provide a more appropriate model for the data than the normal distribution. We therefore need to seek a suitable model for the development over time of a Poisson variable rather than a normal variable. Similarly, there are cases where a linear model fails to represent the behaviour of the data to an adequate extent. For example, if the trend and seasonal terms of a series combine multiplicatively but the disturbance term is additive, a linear model is inappropriate.

In Part II we discuss approximate and exact approaches for handling broad classes of non-Gaussian and nonlinear state space models. Approximate methods include the extended Kalman filter and the more recently developed unscented Kalman filter. We further show how mode estimates of the state vector can be obtained. Exact treatments become feasible when we adopt simulation-based methods. For example, we develop a unified methodology based on importance sampling for analysing non-Gaussian and nonlinear models. Filtering methods for our class of models have become known as particle filtering and we explore the associated methods in detail. Finally, we discuss Bayesian treatments that are also based on simulation methods. A set of illustrations are provided to give the reader a flavour of what can be realised in practice.

This page intentionally left blank

9 Special cases of nonlinear and non-Gaussian models

9.1 Introduction

In this chapter we shall discuss the classes of non-Gaussian and nonlinear models that we shall consider in Part II of this book; we leave aside the analysis of observations generated by these models until later chapters.

A general form of the *nonlinear non-Gaussian state space model* is given by

$$y_t \sim p(y_t|\alpha_t), \qquad \alpha_{t+1} \sim p(\alpha_{t+1}|\alpha_t), \qquad \alpha_1 \sim p(\alpha_1), \tag{9.1}$$

for $t = 1, \dots, n$. We shall assume throughout that

$$p(Y_n|\alpha) = \prod_{t=1}^{n} p(y_t|\alpha_t), \qquad p(\alpha) = p(\alpha_1) \prod_{t=1}^{n-1} p(\alpha_{t+1}|\alpha_t), \tag{9.2}$$

where $Y_n = (y_1', \dots, y_n')'$ and $\alpha = (\alpha_1', \dots, \alpha_n')'$. The observation density $p(y_t|\alpha_t)$ implies a relationship between the observation vector y_t and state vector α_t. The state update density $p(\alpha_{t+1}|\alpha_t)$ implies a relationship between the state vector of the next period α_{t+1} and the state of the current period α_t. If relationships in both $p(y_t|\alpha_t)$ and $p(\alpha_{t+1}|\alpha_t)$ are linear we say that the model is a *linear non-Gaussian state space model*. If all densities $p(y_t|\alpha_t)$, $p(\alpha_{t+1}|\alpha_t)$ and $p(\alpha_1)$ are Gaussian but at least one relationship in $p(y_t|\alpha_t)$ or $p(\alpha_{t+1}|\alpha_t)$ is nonlinear we say that the model is a *nonlinear Gaussian state space model*. In Section 9.2 we consider an important special form of the general linear non-Gaussian model and in Sections 9.3, 9.4, 9.5 and 9.6 we consider special cases of some subclasses of models of interest, namely exponential family models, heavy-tailed models, stochastic volatility model and other financial models. In Section 9.7 we describe some classes of nonlinear models of interest.

9.2 Models with a linear Gaussian signal

The multivariate model with a linear Gaussian signal that we consider here has a similar state space structure to (3.1) in the sense that observational vectors y_t are determined by a relation of the form

$$p\left(y_t|\alpha_1, \dots, \alpha_t, y_1, \dots, y_{t-1}\right) = p(y_t|Z_t\alpha_t), \tag{9.3}$$

where the state vector α_t is determined independently of previous observations by the relation

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \qquad \eta_t \sim \mathrm{N}(0, Q_t), \tag{9.4}$$

with the disturbances η_t being serially independent, for $t = 1, \ldots, n$. We define

$$\theta_t = Z_t\alpha_t, \tag{9.5}$$

and refer to θ_t as the *signal.* The density $p\,(y_t|\theta_t)$ can be non-Gaussian, nonlinear or both. In the case $p(y_t|\theta_t)$ is normal and θ_t is linear in y_t, the model reduces to the linear Gaussian model (3.1). While we begin by considering a general form for $p\,(y_t|\theta_t)$, we shall pay particular attention to three special cases:

1. Observations which come from exponential family distributions with densities of the form

$$p\,(y_t|\theta_t) = \exp[y_t'\theta_t - b_t(\theta_t) + c_t(y_t)], \qquad -\infty < \theta_t < \infty, \tag{9.6}$$

   where $b_t(\theta_t)$ is twice differentiable and $c_t(y_t)$ is a function of y_t only;
2. Observations generated by the relation

$$y_t = \theta_t + \varepsilon_t, \qquad \varepsilon_t \sim p(\varepsilon_t), \tag{9.7}$$

   where the ε_t's are non-Gaussian and serially independent.
3. Observations generated by a fixed mean but a stochastically evolving variance over time as in

$$y_t = \mu + \exp(\frac{1}{2}\theta_t)\varepsilon_t, \qquad \varepsilon_t \sim p(\varepsilon_t), \tag{9.8}$$

   where μ is the mean and the ε_t's are not necessarily Gaussian.

The model (9.6) together with (9.4) and (9.5) where η_t is assumed to be Gaussian was introduced by West, Harrison and Migon (1985) under the name the *dynamic generalised linear model.* The origin of this name is that in the treatment of non-time series data, the model (9.6), where θ_t does not depend on t, is called a *generalised linear model.* In this context θ_t is called the *link function*; for a treatment of generalised linear models see McCullagh and Nelder (1989). Further development of the West, Harrison and Migon model is described in West and Harrison (1997). Smith (1979, 1981) and Harvey and Fernandes (1989) gave an exact treatment for the special case of a Poisson observation with mean modelled as a local level model; their approach, however, does not lend itself to generalisation. In Section 9.3 we discuss a range of interesting and often used densities that are part of the exponential family.

The model (9.7) is similar to the linear Gaussian state space model of Part I with the difference that at least one element of the observation disturbance vector ε_t is non-Gaussian. The typical example is when the observations are contaminated by outliers. In such cases the Gaussian density is not sufficiently strong in the tails of the distribution. The Student's t distribution and the mixture of normals may be more appropriate for $p(\varepsilon_t)$. The details are provided in Section 9.4.

The model (9.8) is known as the *stochastic volatility* (SV) model and is regarded as the parameter-driven counterpart of the observation-driven *generalised autoregressive conditionally heteroscedasticity* (GARCH) model that is described by Engle (1982) and Bollerslev (1986). For a collection of articles that represents the developments of the SV model we refer to Shephard (2005).

9.3 Exponential family models

For model (9.6), let

$$\dot{b}_t(\theta_t) = \frac{\partial b_t(\theta_t)}{\partial \theta_t} \quad \text{and} \quad \ddot{b}_t(\theta_t) = \frac{\partial^2 b_t(\theta_t)}{\partial \theta_t \partial \theta_t'}. \tag{9.9}$$

For brevity, we will write $\dot{b}_t(\theta_t)$ as $\dot{b}_t$ and $\ddot{b}_t(\theta_t)$ as $\ddot{b}_t$ in situations where it is unnecessary to emphasise the dependence on θ_t. Assuming that the relevant regularity conditions are satisfied, it follows immediately by differentiating the relation $\int p(y_t|\theta_t)\, \mathrm{d}y_t = 1$ once and twice that

$$\mathrm{E}(y_t) = \dot{b}_t \quad \text{and} \quad \mathrm{Var}(y_t) = \ddot{b}_t.$$

Consequently $\ddot{b}_t$ must be positive definite for nondegenerate models. The standard results

$$\begin{aligned} \mathrm{E}\left[\frac{\partial \log p(y_t|\theta_t)}{\partial \theta_t}\right] &= 0, \\ \mathrm{E}\left[\frac{\partial^2 \log p(y_t|\theta_t)}{\partial \theta_t \partial \theta_t'}\right] + \mathrm{E}\left[\frac{\partial \log p(y_t|\theta_t)}{\partial \theta_t}\frac{\partial \log p(y_t|\theta_t)}{\partial \theta_t'}\right] &= 0, \end{aligned} \tag{9.10}$$

are obtained directly from (9.6).

9.3.1 Poisson density

For our first example of an exponential family distribution, suppose that the univariate observation y_t comes from a Poisson distribution with mean μ_t. For example, y_t could be the number of road accidents in a particular area during the month. Observations of this kind are called *count data*.

The logdensity of y_t is

$$\log p(y_t|\mu_t) = y_t \log \mu_t - \mu_t - \log(y_t!). \tag{9.11}$$

Comparing (9.11) with (9.6) we see that we need to take $\theta_t = \log \mu_t$ and $b_t = \exp \theta_t$ with $\theta_t = Z_t\alpha_t$, so the density of y_t given the signal θ_t is

$$p(y_t|\theta_t) = \exp[y_t\theta_t - \exp \theta_t - \log y_t!], \qquad t = 1, \dots, n. \tag{9.12}$$

It follows that the mean $\dot{b}_t = \exp \theta_t = \mu_t$ equals the variance $\ddot{b}_t = \mu_t$. Mostly we will assume that η_t in (9.4) is generated from a Gaussian distribution but all or some elements of η_t may come from other continuous distributions.

9.3.2 Binary density

An observation y_t has a binary distribution if the probability that $y_t = 1$ has a specified probability, say π_t, and the probability that $y_t = 0$ is $1 - \pi_t$. For example, we could score 1 if Cambridge won the Boat Race in a particular year and 0 if Oxford won.

Thus the density of y_t is

$$p(y_t|\pi_t) = \pi_t^{y_t}(1 - \pi_t)^{1-y_t}, \qquad y_t = 1, 0, \tag{9.13}$$

so we have

$$\log p(y_t|\pi_t) = y_t[\log \pi_t - \log(1 - \pi_t)] + \log(1 - \pi_t). \tag{9.14}$$

To put this in form (9.6) we take $\theta_t = \log[\pi_t/(1 - \pi_t)]$ and $b_t(\theta_t) = \log(1 + e^{\theta_t})$, and the density of y_t given the signal θ_t is

$$p(y_t|\theta_t) = \exp[y_t\theta_t - \log(1 + \exp \theta_t)], \tag{9.15}$$

for which $c_t = 0$. It follows that mean and variance are given by

$$\dot{b}_t = \frac{\exp \theta_t}{1 + \exp \theta_t} = \pi_t, \qquad \ddot{b}_t = \frac{\exp \theta_t}{(1 + \exp \theta_t)^2} = \pi_t(1 - \pi_t),$$

as is well-known.

9.3.3 Binomial density

Observation y_t has a binomial distribution if it is equal to the number of successes in k_t independent trials with a given probability of success, say π_t. As in the binary case we have

$$\log p(y_t|\pi_t) = y_t[\log \pi_t - \log(1 - \pi_t)] + k_t \log(1 - \pi_t) + \log \begin{pmatrix} k_t \\ y_t \end{pmatrix}, \tag{9.16}$$

with $y_t = 0, \dots, k_t$. We therefore take $\theta_t = \log[\pi_t/(1-\pi_t)]$ and $b_t(\theta_t) = k_t \log(1+\exp\theta_t)$ giving for the density of y_t in form (9.6),

$$p(y_t|\theta_t) = \exp\left[y_t\theta_t - k_t\log(1+\exp\theta_t) + \log\begin{pmatrix} k_t \\ y_t \end{pmatrix}\right]. \tag{9.17}$$

9.3.4 Negative binomial density

There are various ways of defining the negative binomial density; we consider the case where y_t is the number of independent trials, each with a given probability π_t of success, that are needed to reach a specified number k_t of successes. The density of y_t is

$$p(y_t|\pi_t) = \begin{pmatrix} k_t - 1 \\ y_t - 1 \end{pmatrix} \pi_t^{k_t}(1-\pi_t)^{y_t-k_t}, \qquad y_t = k_t, k_{t+1}, \dots, \tag{9.18}$$

and the logdensity is

$$\log p(y_t|\pi_t) = y_t\log(1-\pi_t) + k_t[\log\pi_t - \log(1-\pi_t)] + \log\begin{pmatrix} k_t - 1 \\ y_t - 1 \end{pmatrix}. \tag{9.19}$$

We take $\theta_t = \log(1-\pi_t)$ and $b_t(\theta_t) = k_t[\theta_t - \log(1-\exp\theta_t)]$ so the density in the form (9.6) is

$$p(y_t|\theta_t) = \exp\left[y_t\theta_t - k_t\{\theta_t - \log(1-\exp\theta_t)\} + \log\begin{pmatrix} k_t - 1 \\ y_t - 1 \end{pmatrix}\right]. \tag{9.20}$$

Since in nontrivial cases $1-\pi_t < 1$ we must have $\theta_t < 0$ which implies that we cannot use the relation $\theta_t = Z_t\alpha_t$ since $Z_t\alpha_t$ can be negative. A way around the difficulty is to take $\theta_t = -\exp\theta_t^*$ where $\theta_t^* = Z_t\alpha_t$. The mean $\mathrm{E}(y_t)$ is given by

$$\dot{b}_t = k_t\left[1 + \frac{\exp\theta_t}{1-\exp\theta_t}\right] = k_t\left[1 + \frac{1-\pi_t}{\pi_t}\right] = \frac{k_t}{\pi_t},$$

as is well-known.

9.3.5 Multinomial density

Suppose that we have $h > 2$ cells for which the probability of falling in the ith cell is π_{it} and suppose also that in k_t independent trials the number observed in the ith cell is y_{it} for $i = 1, \dots, h$. For example, monthly opinion polls of voting preference: Labour, Conservative, Liberal Democrat, others.

Let $y_t = (y_{1t}, \dots, y_{h-1,t})'$ and $\pi_t = (\pi_{1t}, \dots, \pi_{h-1,t})'$ with $\sum_{j=1}^{h-1}\pi_{jt} < 1$.

Then y_t is multinomial with logdensity

$$\log p(y_t|\pi_t) = \sum_{i=1}^{h-1} y_{it} \left[\log \pi_{it} - \log \left(1 - \sum_{j=1}^{h-1} \pi_{jt} \right) \right] + k_t \log \left(1 - \sum_{j=1}^{h-1} \pi_{jt} \right) + \log C_t, \tag{9.21}$$

for $0 \leq \sum_{i=1}^{h-1} y_{it} \leq k_t$ where

$$C_t = k_t!/ \left[\prod_{i=1}^{h-1} y_{it}! \left(k_t - \sum_{j=1}^{h-1} y_{jt} \right)! \right].$$

We therefore take $\theta_t = (\theta_{1t}, \ldots, \theta_{h-1,t})'$ where $\theta_{it} = \log[\pi_{it}/(1 - \sum_{j=1}^{h-1} \pi_{jt})]$, and

$$b_t(\theta_t) = k_t \log \left(1 + \sum_{i=1}^{h-1} \exp \theta_{it} \right),$$

so the density of y_t in form (9.6) is

$$p(y_t|\theta_t) = \exp \left[y_t'\theta_t - k_t \log \left(1 + \sum_{i=1}^{h-1} \exp \theta_{it} \right) \right] \times C_t. \tag{9.22}$$

9.3.6 Multivariate extensions

Multivariate generalisations of discrete distributions in the exponential family class are usually not straightforward extensions of their univariate counterparts. We therefore do not consider such generalisations here. However, it is relatively straightforward to have a panel of variables which are independent of each other at time t, conditional on signal θ_t. Denote the ith variable in a panel of discrete time series by y_{it} for $i = 1, \ldots, p$. Models with densities of the form

$$p(y_t|\theta_t) = \prod_{i=1}^{p} p_i(y_{it}|\theta_t),$$

where $p_i(y_{it}|\theta_t)$ refers to an univariate density, possibly in the class of the exponential family. Each density p_i can be different and can be mixed with continuous densities. The variables in the panel only share the time series property implied by θ_t. The vector dimension of θ_t can be different from $p \times 1$. In typical cases of interest, the dimension of θ_t can be less than p. In this case, we effectively obtain a nonlinear non-Gaussian version of the dynamic factor model discussed in Section 3.7.

9.4 Heavy-tailed distributions

9.4.1 *t*-distribution

A common way to introduce error terms into a model with heavier tails than those of the normal distribution is to use Student's t. We therefore consider modelling ε_t of (9.7) by the t-distribution with logdensity

$$\log p(\varepsilon_t) = \log a(\nu) + \frac{1}{2}\log\lambda - \frac{\nu+1}{2}\log\left(1+\lambda\varepsilon_t^2\right), \tag{9.23}$$

where ν is the number of degrees of freedom and

$$a(\nu) = \frac{\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \quad \lambda^{-1} = (\nu-2)\sigma_\varepsilon^2, \quad \sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon_t), \quad \nu > 2, \quad t = 1,\ldots,n.$$

The mean of ε_t is zero and the variance is σ_ε^2 for any ν degrees of freedom which need not be an integer. The quantities ν and σ_ε^2 can be permitted to vary over time, in which case λ also varies over time.

9.4.2 Mixture of normals

A second common way to represent error terms with tails that are heavier than those of the normal distribution is to use a mixture of normals with density

$$p(\varepsilon_t) = \frac{\lambda^*}{\left(2\pi\sigma_\varepsilon^2\right)^{\frac{1}{2}}}\exp\left(\frac{-\varepsilon_t^2}{2\sigma_\varepsilon^2}\right) + \frac{1-\lambda^*}{\left(2\pi\chi\sigma_\varepsilon^2\right)^{\frac{1}{2}}}\exp\left(\frac{-\varepsilon_t^2}{2\chi\sigma_\varepsilon^2}\right), \tag{9.24}$$

where λ^* is near to one, say 0.95 or 0.99, and χ is large, say from 10 to 100. This is a realistic model for situations when outliers are present, since we can think of the first normal density of (9.24) as the basic error density which applies $100\lambda^*$ per cent of the time, and the second normal density of (9.24) as representing the density of the outliers. Of course, λ^* and χ can be made to depend on t if appropriate. The investigator can assign values to λ^* and χ but they can also be estimated when the sample is large enough.

9.4.3 General error distribution

A third heavy-tailed distribution that is sometimes used is the general error distribution with density

$$p(\varepsilon_t) = \frac{w(\ell)}{\sigma_\varepsilon}\exp\left[-c(\ell)\left|\frac{\varepsilon_t}{\sigma_\varepsilon}\right|^\ell\right], \qquad 1 < \ell < 2, \tag{9.25}$$

where

$$w(\ell) = \frac{2[\Gamma(3\ell/4)]^{\frac{1}{2}}}{\ell[\Gamma(\ell/4)]^{\frac{3}{2}}}, \qquad c(\ell) = \left[\frac{\Gamma(3\ell/4)}{\Gamma(\ell/4)}\right]^{\frac{\ell}{2}}.$$

Some details about this distribution are given by Box and Tiao (1973, §3.2.1), from which it follows that $\mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2$ for all ℓ.

9.5 Stochastic volatility models

In the standard state space model (3.1) the variance of the observational error ε_t is assumed to be constant over time. In the analysis of financial time series, such as daily fluctuations in stock prices and exchange rates, return series will usually be approximately serially uncorrelated. Return series may not be serially independent, however, because of serial dependence in the variance. It is often found that the observational error variance is subject to substantial variability over time. This phenomenon is referred to as *volatility clustering.* An allowance for this variability in models for such series may be achieved via the *stochastic volatility* (SV) *model.* The SV model has a strong theoretical foundation in the financial theory on option pricing based on the work of the economists Black and Scholes; for a discussion see Taylor (1986). Further, the SV model has a strong connection with the state space approach as will become apparent below.

Denote the first (daily) differences of a particular series of asset log prices by y_t. Financial time series are often constructed by first differencing log prices of some portfolio of stocks, bonds, foreign currencies, etc. A basic SV model for y_t is given by

$$y_t = \mu + \sigma \exp\left(\frac{1}{2}\theta_t\right)\varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0,1), \qquad t = 1,\dots,n, \tag{9.26}$$

where the mean μ and the average standard deviation σ are assumed fixed and unknown. The signal θ_t is regarded as the unobserved log-volatility and it can be modelled in the usual way by $\theta_t = Z_t\alpha_t$ where α_t is generated by (9.4). In standard cases θ_t is modelled as an AR(1) process with Gaussian disturbances, that is $\theta_t = \alpha_t$ where α_t is the state process

$$\alpha_{t+1} = \phi\alpha_t + \eta_t, \qquad \eta_t \sim \mathrm{N}\big(0,\sigma_\eta^2\big), \qquad 0 < \phi < 1, \tag{9.27}$$

for $t = 1,\dots,n$ and with $\alpha_1 \sim \mathrm{N}[0, \sigma_\eta^2\,/\,(1-\phi^2)]$. In other cases, the generality of the state equation (9.4) for α_t can be fully exploited. The model can be regarded as the discrete time analogue of the continuous time model used in papers on option pricing, such as Hull and White (1987). Since the model is Gaussian, $\mathrm{E}(y_t|\theta_t) = \mu$ and $\mathrm{Var}(y_t|\theta_t) = \sigma^2\exp(\theta_t)$, it follows that the logdensity of the SV model (9.26) is given by

$$\log p(y_t|\theta_t) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\theta_t - \frac{1}{2\sigma^2}(y_t-\mu)^2\exp(-\theta_t).$$

Further statistical properties of y_t are easy to determine. However, the model is not linear so the techniques described in Part I of this book cannot provide an exact solution for statistical analysis. For a review of work and developments of the SV model see Shephard (1996), Ghysels, Harvey and Renault (1996) and Shephard (2005).

Parameter estimation for the SV models based on maximum likelihood has been considered elsewhere as a difficult problem. Linear Gaussian techniques only offer approximate maximum likelihood estimates of the parameters and can only be applied to the basic SV model (9.26). The techniques we develop in the following chapters of this book, however, provide analyses of SV models based on simulation methods which can be made as accurate as is required.

Various extensions of the SV model can be considered. In the remainder of this section we discuss a number of such extensions together with related models.

9.5.1 Multiple volatility factors

In empirical work it has been observed that volatility often exhibits long-range dependence; see for example Andersen, Bollerslev, Diebold and Labys (2003). Ideally, log-volatility θ_t is modelled by a fractionally integrated process, for example; see Granger and Joyeau (1980). Inference for the SV model (9.26) with a long memory process for θ_t is often based on the spectral likelihood function; see, for example, Breidt, Crato and de Lima (1998) and Ray and Tsay (2000). Exact maximum likelihood methods have recently been considered by Brockwell (2007). In our framework, we can approximate the long-range dependence in the log-volatility θ_t by considering it as a sum of independent autoregressive factors, that is

$$\theta_t = \sum_{i=1}^{q} \theta_{it},$$

where each θ_{it} represents an independent process as in (9.27). The most commonly used specification is the two factor model ($q = 2$), where one factor can be associated with the long-run dependence and the other with the short-run dependence; see the discussion in Durham and Gallant (2002).

9.5.2 Regression and fixed effects

The basic SV model (9.26) captures only the salient features of changing volatility in financial series over time. The model becomes more precise when the mean of y_t is modelled by incorporating explanatory variables. For example, the SV model may be formulated as

$$b(L)y_t = \mu + c(L)'x_t + \sigma \exp\left(\frac{1}{2}\theta_t\right)\varepsilon_t,$$

where L is the lag operator defined so that $L^j z_t = z_{t-j}$ for $z_t = y_t, x_t$ and where $b(L) = 1 - b_1 L - \cdots - b_{p^*} L^{p^*}$ is a scalar lag polynomial of order p^*; the column vector polynomial $c(L) = c_0 + c_1 L_1 + \cdots + c_{k^*} L^{k^*}$ contains $k^* + 1$ vectors of coeffients and x_t is a vector of exogenous explanatory variables. Note that the lagged value y_{t-j}, for $j = 1, \ldots, p^*$, can be considered as an explanatory variable to be added to exogenous explanatory variables. An illustration is provided by

Tsiakas (2006) who introduce dummy effects to account for a seasonal pattern in the volatility. Koopman, Jungbacker and Hol (2005) consider a regression variable that contains information on the unobserved log-volatility process. Such regression effects can be incorporated into the SV model by letting the signal θ_t depend on explanatory variables.

9.5.3 Heavy-tailed disturbances

The Gaussian density $p(\varepsilon_t)$ in the SV model can be replaced by a density with heavier tails such as the t-distribution. This extension is often appropriate because many empirical studies find outlying returns (mostly negative but also positive) due to unexpected jumps or downfalls in asset prices caused by changes in economic conditions or turmoil in financial markets. Key examples are the 'black Monday' crash in October 1987 and the world-wide banking crisis in the second half of 2008. The resulting excess kurtosis found in time series of financial returns can be modelled by having a standardised Student's t distribution for ε_t in (9.26) and its density given by (9.23). The dynamic properties of logvolatility and the thickness of tails are modelled separately as a result. Examples of this approach can be found in Fridman and Harris (1998), Liesenfeld and Jung (2000) and Lee and Koopman (2004).

9.5.4 Additive noise

In the empirical finance literature it is widely recognised that financial prices or returns which are observed at very short time intervals are subject to noise due to discrete observed prices, market regulations and market imperfections. The last source is related to strategic trading behaviour and is commonly caused by differences in the amount of information that traders have about the market. This phenomenon is collectively referred to as *market micro-structure effects* which become more and more apparent as prices are observed at smaller and smaller time intervals; see Campbell, Lo and MacKinlay (1997) and the references therein for a further discussion.

The basic SV model assumes that financial returns only have one source of error. In case the returns are observed at a higher frequency, the SV model should be extended with an additive noise component to account for market microstructure. The additive noise can be represented by a Gaussian disturbance term with constant variance. More specifically, we have

$$y_t = \mu + \sigma \exp\left(\frac{1}{2}\theta_t\right)\varepsilon_t + \zeta_t, \qquad \varepsilon_t \sim \mathrm{N}(0,1), \quad \zeta_t \sim \mathrm{N}(0,\sigma_\zeta^2), \tag{9.28}$$

where all disturbances are serially uncorrelated. The disturbance term ζ_t represents market microstructure effects in the returns. This model has been considered by Jungbacker and Koopman (2005).

9.5.5 Leverage effects

Another characteristic of financial time series is the phenomenon of *leverage*. The volatility of financial markets may adapt differently to positive and negative shocks. It is often observed that while markets might remain more or less stable when large positive earnings have been achieved but when huge losses have to be digested, markets become more unpredictable in the periods ahead. The start of the banking crisis in September 2008 is a clear illustration of the leverage effect. In the seminal paper of Black (1976), the phenomenon of leverage is described. In terms of the SV model (9.26) and (9.27), the leverage effect occurs if a negative return ($\varepsilon_t < 0$) increases the volatility ($\eta_t > 0$) more than a positive return ($\varepsilon_t > 0$) of the same magnitude decreases it ($\eta_t < 0$). The leverage effect is incorporated in the SV model by allowing correlation between the disturbances of the state and the observation equation; see Yu (2005) for a detailed discussion. In the case of our basic SV model (9.26) and (9.27), we achieve this by having

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim \mathrm{N}\left(0, \begin{bmatrix} 1 & \sigma_\eta \rho \\ \sigma_\eta \rho & \sigma_\eta^2 \end{bmatrix}\right),$$

for $t = 1, \ldots, n$. The correlation coefficient ρ is typically negative, implying that negative shocks in the return are accompanied by positive shocks in the volatility and vice versa.

The nonlinear state space formulation of the SV model with leverage requires both θ_t and η_t in the state vector α_t to account for the nonlinear relationship. For this purpose, a more convenient specification of the SV model with leverage is proposed by Jungbacker and Koopman (2007) where the model is reformulated as

$$y_t = \sigma \exp(\frac{1}{2} h_t^*) \left\{ \varepsilon_t^* + \mathrm{sign}(\rho) \xi_{2t} \right\}, \qquad \varepsilon_t^* \sim \mathrm{N}(0, 1 - |\rho|),$$

where

$$h_{t+1}^* = \phi h_t^* + \sigma_\xi \left(\xi_{1,t} + \xi_{2t} \right), \qquad \xi_{1t} \sim \mathrm{N}(0, 1 - |\rho|), \qquad \xi_{2t} \sim \mathrm{N}(0, |\rho|),$$

for $t = 1, \ldots, n$, with $h_1^* \sim N\{0, \sigma_\xi^2 (1 - \phi^2)^{-1}\}$. The disturbances ε_t^*, ξ_{1t} and ξ_{2t} are mutually and serially independent for $t = 1, \ldots, n$. In terms of the general formulation (9.4), we have $\alpha_t = (h_t^* \ , \ \sigma_\xi \xi_{2,t})'$, $\xi_t = \sigma_\xi (\xi_{1,t} \ , \ \xi_{2,t+1})'$ and

$$\theta_t = \alpha_t, \quad \alpha_{t+1} = \begin{bmatrix} \phi & 1 \\ 0 & 0 \end{bmatrix} \alpha_t + \xi_t, \qquad \begin{array}{lcl} \xi_t & \sim & \mathrm{N}\left\{0, \sigma_\xi^2 \mathrm{diag}(1 - |\rho|, |\rho|)\right\}, \\ \alpha_1 & \sim & \mathrm{N}\left\{0, \sigma_\xi^2 \mathrm{diag}([1 - \phi^2]^{-1}, |\rho|)\right\}, \end{array}$$

for $t = 1, \ldots, n$. The observations $y_1, \ldots, y_n$ have the conditional density of the form (9.3) and is given by

$$\log p(y_n | \theta) = \sum_{t=1}^{n} \log p(y_t | \theta_t),$$

where

$$\log p(y_t|\theta_t) = c - \frac{1}{2}h_t^* - \frac{1}{2}\sigma^{-2}\exp(-h_t^*)(1-|\rho|)^{-1}\{y_t - \sigma\exp(\frac{1}{2}h_t^*)\text{sign}(\rho)\xi_{2,t}\}^2,$$

for $t = 1, \ldots, n$ where c is some constant.

9.5.6 Stochastic volatility in mean

As investors require a larger expected return if the risk is large, it seems reasonable to expect a positive relationship between volatility and returns. Empirical evidence however points to a negative influence of volatility on returns; see, for example, French, Schwert and Stambaugh (1987). This effect can be explained by assuming a positive relationship between expected return and *ex-ante* volatility. Koopman and Hol-Uspensky (2002) proposed capturing this so-called volatility feedback effect by including volatility as a regression effect in the mean function. Such a model is labelled as the *SV in Mean* (SVM) model and its simplest form is given by

$$y_t = \mu + d\exp(\theta_t) + \sigma\exp\left(\frac{1}{2}\theta_t\right)\varepsilon_t,$$

where d is the risk premium coefficient which is fixed and unknown. Other forms of the SVM model may also be considered but this one is particularly convenient.

9.5.7 Multivariate SV models

Consider a $p \times 1$ vector of differenced series of asset log prices $y_t = (y_{1t}, \ldots, y_{pt})'$ with constant mean $\mu = (\mu_1, \ldots, \mu_p)'$ and stochastic time-varying variance matrix V_t. The basic version of the multivariate stochastic volatility model can be given by

$$y_t = \mu + \varepsilon_t, \qquad \varepsilon_t \sim \text{N}(0, V_t), \qquad t = 1, \ldots, n, \tag{9.29}$$

where time-varying variance matrix V_t is a function of the scalar or vector signal θ_t as given by (9.5), that is $V_t = V_t(\theta_t)$. The model (9.29) implies that $y_t|\theta_t \sim \text{N}(\mu, V_t)$, for $t = 1, \ldots, n$. We discuss three possible specifications of the variance matrix $V_t(\theta_t)$ below. Other multivariate generalisations of the univariate stochastic volatility model can be considered as well. A more extensive discussion of the multivariate SV model is presented in Asai and McAleer (2005). A treatment of the three multivariate models below is given by Jungbacker and Koopman (2006).

The first multivariate SV model is based on a single time-varying factor. We can take the variance matrix V_t as a constant matrix and scale it by a stochastically time-varying scalar θ_t, that is

$$V_t = \exp(\theta_t)\Sigma_\varepsilon, \qquad t = 1, \ldots, n. \tag{9.30}$$

This multivariate generalisation of the univariate SV model implies observations y_t with time-varying variances and covariances but with correlations that are constant over time. The conditional density $p(y_t|\theta_t)$ is given by

$$p(y_t|\theta_t) = -\frac{p}{2}\log 2\pi - \frac{p}{2}\theta_t - \frac{1}{2}\log|\Sigma_\varepsilon| - \frac{1}{2}\exp(-\theta_t)s_t, \qquad t = 1,\dots,n,$$

with scalar $s_t = (y_t - \mu)'\Sigma_\varepsilon^{-1}(y_t - \mu)$. The formulation (9.30) was originally proposed by Quintana and West (1987) where they used Bayesian methods for inference. Shephard (1994a) proposed a similar model and referred to it as the local scale model. A further extension for the linear Gaussian state space model where all variances are scaled by the common stochastic scalar $\exp(\theta_t)$ is considered by Koopman and Bos (2004).

The second model has stochastically time-varying variances but constant correlations. We consider a $p \times 1$ vector of log-volatilities $\theta_t = (\theta_{1t},\dots,\theta_{pt})'$. The multivariate extension of the basic SV model (9.26) that is considered by Harvey, Ruiz and Shephard (1994) is given by

$$y_t = \mu + D_t\varepsilon_t, \qquad D_t = \exp\{\frac{1}{2}\text{diag}(\theta_{1t},\dots,\theta_{pt})\}, \qquad \varepsilon_t \sim \text{N}(0,\Sigma_\varepsilon), \quad (9.31)$$

for $t = 1,\dots,n$. The conditional variance matrix of y_t is given by $V_t = D_t\Sigma_\varepsilon D_t$ where the constant matrix Σ_ε is effectively a correlation matrix with unity values on its leading diagonal. The variance matrix V_t is a stochastic function of time but the correlations are constant over time. The conditional density is given by

$$p(y_t|\theta_t) = -\frac{p}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{p}\theta_{it} - \frac{1}{2}\log|\Sigma_\varepsilon| - \frac{1}{2}s_t'\Sigma_\varepsilon^{-1}s_t, \qquad t = 1,\dots,n.$$

where $s_t = D_t^{-1}(y_t - \mu)$ is a $p \times 1$ vector with its ith element equal to $s_{it} = \exp(-0.5\theta_{it})(y_{it} - \mu_i)$ for $i = 1,\dots,p$.

The third model has time-varying variances and correlations. It is based on model (9.29) with the variance matrix decomposed as $V_t = CD_t^2C'$ where matrix C is a lower unity triangular matrix and D_t is specified as in (9.31). The variance matrix is effectively subject to a Cholesky decomposition with a time-varying D_t^2. In this specification both the variances and the correlations implied by V_t are time-varying. The resulting model belongs to a class of multivariate SV models that were originally proposed by Shephard (1996) and further extended and analysed by Aguilar and West (2000) and Chib, Nardari and Shephard (2006). The general class of this model allows for a number of $r < p$ 'volatility factors' where θ_t is an $r \times 1$ vector and the $p \times r$ matrix C contains loading factors and includes an additive disturbance vector with constant variances in (9.29).

9.5.8 Generalised autoregressive conditional heteroscedasticity

The generalised autoregressive conditional heteroscedasticity (GARCH) model, a special case of which was introduced by Engle (1982) and is known as the ARCH model, is a widely discussed model in the financial and econometrics literature. A simplified version of the GARCH(1, 1) model is given by

$$\begin{aligned} y_t &= \sigma_t \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0,1), \\ \sigma_{t+1}^2 &= \alpha^* y_t^2 + \beta^* \sigma_t^2, \end{aligned} \tag{9.32}$$

where the parameters to be estimated are α^* and β^*. For a review of the GARCH model and its extensions see Bollerslev, Engle and Nelson (1994).

It is shown by Barndorff-Nielsen and Shephard (2001) that recursion (9.32) is equivalent to the steady state Kalman filter for a particular representation of the SV model. Consider the model

$$\begin{aligned} y_t &= \sigma_t \varepsilon_t, & \varepsilon_t &\sim \mathrm{N}(0,1), \\ \sigma_{t+1}^2 &= \phi \sigma_t^2 + \eta_t, & \eta_t &> 0, \end{aligned} \tag{9.33}$$

for $t = 1, \ldots, n$, where disturbances ε_t and η_t are serially and mutually independently distributed. Possible distributions for η_t are the gamma, inverse gamma or inverse Gaussian distributions. We can write the model in its squared form as follows

$$y_t^2 = \sigma_t^2 + u_t, \qquad u_t = \sigma_t^2 \left(\varepsilon_t^2 - 1 \right),$$

which is in a linear state space form with $\mathrm{E}(u_t) = 0$. The Kalman filter provides the minimum mean squared error estimate a_t of σ_t^2. When in steady state, the Kalman update equation for a_{t+1} can be represented as the GARCH$(1,1)$ recursion

$$a_{t+1} = \alpha^* y_t^2 + \beta^* a_t,$$

with

$$\alpha^* = \phi \frac{\bar{P}}{\bar{P}+1}, \qquad \beta^* = \phi \frac{1}{\bar{P}+1},$$

where $\bar{P}$ is the steady state value for P_t of the Kalman filter which we have defined for the local level model in Section 2.11 and for the general linear model in Subsection 4.3.4. We note that $\alpha^* + \beta^* = \phi$.

9.6 Other financial models

9.6.1 Durations: exponential distribution

Consider a series of transactions in a stock market in which the tth transaction x_t is time-stamped by the time τ_t at which it took place. When studying the behaviour of traders in the market, attention may be focused on the duration between successive transactions, that is $y_t = \Delta \tau_t = \tau_t - \tau_{t-1}$. The duration y_t with mean μ_t can be modelled by a simple exponential density given by

$$p(y_t | \mu_t) = \frac{1}{\mu_t} \exp(-y_t / \mu_t), \qquad y_t, \mu_t > 0. \tag{9.34}$$

This density is a special case of the exponential family of densities and to put it in the form (9.6) we define

$$\theta_t = -\frac{1}{\mu_t} \quad \text{and} \quad b_t(\theta_t) = \log \mu_t = -\log(-\theta_t),$$

so we obtain

$$\log p(y_t|\theta_t) = y_t\theta_t + \log(-\theta_t). \tag{9.35}$$

Since $\dot{b}_t = -\theta_t^{-1} = \mu_t$ we confirm that μ_t is the mean of y_t, as is obvious from (9.34). The mean is restricted to be positive and so we model $\theta_t^* = \log(\mu_t)$ rather than μ_t directly. The durations in financial markets are typically short at the opening and closing of the daily market hours due to heavy trading in these periods. The time stamp τ_t is therefore often used as an explanatory variable in the mean function of durations and in order to smooth out the huge variations of this effect, a cubic spline is used. A simple durations model which allows for the daily seasonal is then given by

$$\begin{aligned} \theta_t &= \gamma(\tau_{t-1}) + \psi_t, \\ \psi_t &= \rho\psi_{t-1} + \chi_t, \qquad \chi_t \sim \mathrm{N}\big(0, \sigma_\chi^2\big), \end{aligned}$$

where $\gamma(\cdot)$ is the cubic spline function and χ_t is serially uncorrelated. Such models can be regarded as state space counterparts of the influential *autoregressive conditional duration* (ACD) model of Engle and Russell (1998).

9.6.2 Trade frequencies: Poisson distribution

Another way of analysing market activity is to divide the daily market trading period into intervals of one or five minutes and record the number of transactions in each interval. The counts in each interval can be modelled by a Poisson density for which the details are given in Subsection 9.3.1. Such a model would be a basic discrete version of what Rydberg and Shephard (2003) have labelled as BIN models.

9.6.3 Credit risk models

A firm can obtain its credit rating from commercial agencies such as Moody's and Standard & Poors. A migration from one rating class to another is indicative of the performance of the firm but also of the economic conditions under which the firm and its trading and financial partners operate. Credit risk indicators aim to provide an insight into the overall direction of the rating migrations for an industry or for the economy as a whole. Such indicators are of key interest to financial regulators and economic policy makers. The construction of a credit risk indicator from a database of credit ratings of firms is a challenging task since we need to consider a large database with the rating history for each firm. In a credit risk model, the rating itself can be regarded as a stochastic variable but also the duration at which the firm keeps it rating before it enters into a different rating category. Since the ratings are measured in classes such as AAA, AA, A, BBB, and so on, the rating variable is inherently a non-Gaussian variable. Koopman, Lucas and Monteiro (2008) accommodate the stylised properties of credit rating migrations by an intensity-based duration model for different types of migrations and driven by a common signal θ_t with the possibility of including explanatory variables as well. The common signal represents the overall credit risk indicator. This modelling framework can be cast in the general class of models discussed in Section 9.2.

A simplification of the analysis can be obtained by compressing the available data into numbers of upgrades, downgrades and defaults in each week or month for different categories. Such counts are usually small; especially when we focus on specific groups of firms (manufacturing, banking, transport, etc.). Hence we treat these counts as coming from a binomial distribution. This is the approach taken by Koopman and Lucas (2008) who consider a panel of N time series of counts y_{ijt} where index i refers to a transition type (e.g. number of downgrades of firms in higher credit rating classes, downgrades of firms that have lower ratings, upgrades, defaults), index j is for a group of firms and index t is for the time period. The counts can be regarded a the number of 'successes' in k_{ijt} independent trials (number of firms in the appropriate group at time t) with a given probability of success, say π_{ijt}. A parsimonious model specification for probability π_{ijt} is given by

$$\pi_{ijt} = \frac{\exp\theta^*_{ijt}}{1+\exp\theta^*_{ijt}}, \qquad \theta^*_{ijt} = \mu_{ij} + \lambda'_{ij}\theta_t,$$

where θ_t is a signal vector that we can specify as (9.5) and where the scalar coefficients μ_{ij} and the vector coefficients λ_{ij} are treated as unknown fixed parameters that need to be estimated. The constants μ_{ij} and the factor loading vectors λ_{ij} can be pooled into a smaller set of unknown coefficients. The binomial density (9.16) discussed in Subsection 9.3.3 may be appropriate for y_{ijt}. When we further assume that the observations conditional on the dynamic factors are independent of each other, we can formulate the conditional density at time t as the product of the individual densities for all i and j. Hence we have shown that this modelling framework for a credit risk analysis fits naturally in a multivariate extension of the exponential family models that we discussed in Section 9.3. Further extensions with economic and financial variables and with constructed business cycle indicators as explanatory variables are considered by Koopman, Lucas and Schwaab (2011).

9.7 Nonlinear models

In this section we introduce a class of nonlinear models which is obtained from the standard linear Gaussian model (3.1) in a natural way by permitting y_t to depend nonlinearly on α_t in the observation equation and α_{t+1} to depend nonlinearly on α_t in the state equation. Thus we obtain the model

$$y_t = Z_t(\alpha_t) + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, H_t), \tag{9.36}$$

$$\alpha_{t+1} = T_t(\alpha_t) + R_t\eta_t, \qquad \eta_t \sim \mathrm{N}(0, Q_t), \tag{9.37}$$

for $t = 1, \dots, n$, with $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ and where $Z_t(\cdot)$ and $T_t(\cdot)$ are differentiable vector functions of α_t with dimensions p and m respectively. In principle it would be possible to extend this model by permitting ε_t and η_t to be non-Gaussian but we shall not pursue this extension in this book. Models with general forms similar to this were considered by Anderson and Moore (1979).

A simple example of the relation (9.36) is a nonlinear version of the structural time series model in which the trend μ_t and seasonal γ_t combine multiplicatively and the observation error ε_t is additive, giving

$$y_t = \mu_t \gamma_t + \varepsilon_t;$$

a model of this kind has been considered by Shephard (1994b). A related and more general model is proposed by Koopman and Lee (2009) and is based on the specification

$$y_t = \mu_t + \exp(c_0 + c_\mu \mu_t)\gamma_t + \varepsilon_t,$$

where c_0 and c_μ are unknown coefficients. In these nonlinear models, the magnitude of the seasonal fluctuations in the realisations of y_t depend on the trend in the series. This feature of a time series is often encountered and the typical action taken is to transform y_t by taking logs. The models here provide an alternative to this data transformation and become more relevant when observations cannot be transformed by taking logs due to having negative values in the data.

10 Approximate filtering and smoothing

10.1 Introduction

In this chapter we consider approximate filtering and smoothing for data generated by a variety of non-Gaussian and nonlinear models such as those exemplified in Chapter 9. For the purpose of filtering, we assume that new observations y_t come in one at a time and that we wish to estimate functions of the state vector sequentially at each time point t taking account of all the observations up to and including time t. In the case of smoothing, we wish to estimate functions of the state vector for a given set of observations $y_1, \ldots, y_n$. Motivations and illustrations of filtering and smoothing for linear Gaussian models are discussed in Part I of this book. These motivations and illustrations are not intrinsically different in nonlinear non-Gaussian cases but the expressions for filtering and smoothing are not available in an analytical form. We rely on approximations or numerical solutions. Approximations to linear estimation and to Bayesian analysis can be obtained from Lemmas 2, 3 and 4 of Chapter 4.

We begin by considering two approximate filters, *the extended Kalman filter* in Section 10.2 and *the unscented Kalman filter* in Section 10.3. The ideas underlying the two approaches to nonlinear filtering are presented and their numerical performances are compared. Next, in Section 10.4 we consider nonlinear smoothing and show how approximate smoothing recursions can be derived for the two approximate filters. Section 10.5 argues that approximate solutions for filtering and smoothing can also be obtained when the data are transformed in an appropriate way. Two illustrations are presented as examples. In Sections 10.6 and 10.7 we discuss methods for computing the mode estimate of the state and signal vectors. The mode estimates are computed exactly. However, the analysis often focuses on mean, variance and possibly higher moments of the density and therefore provides an approximation in practice. Different treatments for models with heavy-tailed errors are collected and presented in Section 10.8.

10.2 The extended Kalman filter

The *extended Kalman filter* (EKF) is based on the idea of linearising the observation and state equations and then applying the Kalman filter straightforwardly to the resulting linearised model. We start with the following special case of the non-Gaussian nonlinear model (9.36) and (9.37) but where the disturbances are not necessarily normally distributed, that is

$$y_t = Z_t(\alpha_t) + \varepsilon_t, \qquad \alpha_{t+1} = T_t(\alpha_t) + R_t(\alpha_t)\eta_t, \tag{10.1}$$

for $t = 1, \dots, n$, where $Z_t(\alpha_t)$, $T_t(\alpha_t)$ and $R_t(\alpha_t)$ are differentiable functions of α_t and where the random disturbances ε_t and η_t are serially and mutually uncorrelated with mean zero and variance matrices $H_t(\alpha_t)$ and $Q_t(\alpha_t)$, respectively. The initial state vector α_1 is random with mean a_1 and variance matrix P_1, and is uncorrelated with all disturbances.

We adopt the definitions of the predicted and filtered state vectors as used in Chapter 4, that is $a_t = \mathrm{E}(\alpha_t|Y_{t-1})$ and $a_{t|t} = \mathrm{E}(\alpha_t|Y_t)$, respectively. Define

$$\dot{Z}_t = \left.\frac{\partial Z_t(\alpha_t)}{\partial \alpha_t'}\right|_{\alpha_t = a_t}, \qquad \dot{T}_t = \left.\frac{\partial T_t(\alpha_t)}{\partial \alpha_t'}\right|_{\alpha_t = a_{t|t}}, \tag{10.2}$$

where we emphasise that $\dot{Z}_t$ is evaluated at time $t-1$ since a_t is a function of $y_1, \dots, y_{t-1}$ and $\dot{T}_t$ is evaluated at time t since $a_{t|t}$ depends on $y_1, \dots, y_t$. Expanding the matrix functions of (10.1) in Taylor series, based on the appropriate fixed values of a_t and $a_{t|t}$, gives

$$\begin{aligned}
Z_t(\alpha_t) &= Z_t(a_t) + \dot{Z}_t\,(\alpha_t - a_t) + \dots,\\
T_t(\alpha_t) &= T_t(a_{t|t}) + \dot{T}_t\,(\alpha_t - a_{t|t}) + \dots,\\
R_t(\alpha_t) &= R_t(a_{t|t}) + \dots,\\
H_t(\alpha_t) &= H_t(a_t) + \dots,\\
Q_t(\alpha_t) &= Q_t(a_{t|t}) + \dots.
\end{aligned}$$

The matrix function $Z_t(\alpha_t)$ in the observation equation is expanded based on a_t while the matrix functions $T_t(\alpha_t)$ and $R_t(\alpha_t)$ are expanded based on $a_{t|t}$ since y_t is available for the state equation at time $t+1$. Substituting these expressions in (10.1), neglecting higher-order terms and assuming knowledge of a_t and $a_{t|t}$ gives

$$y_t = \dot{Z}_t\alpha_t + d_t + \varepsilon_t, \qquad \alpha_{t+1} = \dot{T}_t\alpha_t + c_t + R_t(a_{t|t})\eta_t, \tag{10.3}$$

where

$$d_t = Z_t(a_t) - \dot{Z}_t a_t, \qquad c_t = T_t(a_{t|t}) - \dot{T}_t a_{t|t},$$

and

$$\varepsilon_t \sim [0, H_t(a_t)], \qquad \eta_t \sim \big[0, Q_t(a_{t|t})\big].$$

Using the minimum variance matrix property of the Kalman filter discussed in Sections 4.2 and 4.3 as justification, we apply the Kalman filter with mean adjustments of Subsection 4.3.3 to the linearised model (10.1). We have

$$\begin{aligned} v_t &= y_t - \dot{Z} a_t - d_t \\ &= y_t - Z_t(a_t), \\ a_{t|t} &= a_t + P_t \dot{Z}_t' F_t^{-1} v_t, \\ a_{t+1} &= \dot{T}_t a_t + K_t v_t + c_t \\ &= \dot{T}_t a_t + K_t v_t + T_t(a_{t|t}) - \dot{T}_t[a_t + P_t \dot{Z}_t' F_t^{-1} v_t] \\ &= T_t(a_{t|t}), \end{aligned}$$

where $F_t = \dot{Z}_t P_t \dot{Z}_t' + H_t(a_t)$ and $K_t = \dot{T}_t P_t \dot{Z}_t' F_t^{-1}$. Putting these formulae together with the other equations from Subsection 4.3.3, we obtain the following recursion for calculating a_{t+1} and P_{t+1} given a_t and P_t,

$$\begin{aligned} v_t &= y_t - Z_t(a_t), & F_t &= \dot{Z}_t P_t \dot{Z}_t' + H_t(a_t), \\ a_{t|t} &= a_t + P_t \dot{Z}_t' F_t^{-1} v_t, & P_{t|t} &= P_t - P_t \dot{Z}_t' F_t^{-1} \dot{Z}_t P_t, \\ a_{t+1} &= T_t(a_{t|t}), & P_{t+1} &= \dot{T}_t P_{t|t} \dot{T}_t' + R_t(a_{t|t}) Q_t(a_{t|t}) R_t(a_{t|t})', \end{aligned} \tag{10.4}$$

for $t = 1, \ldots, n$. This recursion, together with the initial values a_1 and P_1, is called the *extended Kalman filter*. It is essentially given in this form by the equations (2.4) to (2.8) of Anderson and Moore (1979, §8.2). Earlier versions of the extended Kalman filter have appeared in Jazwinski (1970, §8.3). Comparisons of its performance with that of other filters will be presented in Subsection 10.3.4.

The extended Kalman filter is developed to accomodate nonlinear effects in the state space model. In case the densities of the disturbances ε_t and η_t in (10.1) are non-Gaussian, the extended Kalman filter does not change. We assume that the mean of the disturbances are zero and we set $H_t(\alpha_t)$ and $Q_t(\alpha_t)$ equal to the variance matrices of ε_t and η_t, respectively. In case elements of the variance matrices depend on the state vector, α_t is replaced by a_t or $a_{t|t}$.

10.2.1 A multiplicative trend-cycle decomposition

Consider the partly multiplicative model for the univariate observation y_t given by

$$y_t = \mu_t \times c_t + \varepsilon_t,$$

for $t = 1, \ldots, n$, where μ_t is the trend component as modelled by the random walk process $\mu_{t+1} = \mu_t + \xi_t$ and c_t is the unobserved cycle component (3.13) as discussed in Section 3.2. Similar multiplicative models are discussed in Section 9.7. The 3×1 state vector for this model is given by $\alpha_t = (\mu_t\,, c_t\,, c_t^*)'$ where c_t^* is implicitly defined in equation (3.13). The 3×1 disturbance vector is given by $\eta_t = (\xi_t\,, \tilde{\omega}_t\,, \tilde{\omega}_t^*)'$ where the cycle disturbances $\tilde{\omega}_t$ and $\tilde{\omega}_t^*$ are defined in equation (3.13). The three disturbances are independent normal variables. The state equation of (10.1) is linear for the multiplicative trend-cycle model. In particular, we have $T_t(\alpha_t) = T_t \times \alpha_t$, $R_t(\alpha_t) = R_t$ and $Q_t(\alpha_t) = Q_t$ with

$$T_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \rho\cos\lambda_c & \rho\sin\lambda_c \\ 0 & -\rho\sin\lambda_c & \rho\cos\lambda_c \end{bmatrix}, \qquad R_t = I_3, \qquad Q_t = \mathrm{diag}(\sigma_\xi^2\,,\, \sigma_\omega^2\,,\, \sigma_\omega^2),$$

where ρ is a discounting factor with $0 < \rho < 1$, λ_c is the frequency of the cycle c_t, σ_ξ^2 is the variance of disturbance ξ_t and σ_ω^2 is the variance of both disturbances $\tilde{\omega}_t$ and $\tilde{\omega}_t^*$. The multiplicative decomposition is represented by the nonlinear function $Z_t(\alpha_t)$ in the observation equation of (10.1) with the variance matrix $H_t(\alpha_t)$ and are given by

$$Z_t(\alpha_t) = \alpha_{1t}\alpha_{2t}, \qquad H_t(\alpha_t) = \sigma_\varepsilon^2,$$

where α_{jt} is the jth element of α_t and σ_ε^2 is the variance of the irregular or error term ε_t. We notice that $\alpha_{1t} = \mu_t$ and $\alpha_{2t} = c_t$ as required.

To apply the extended Kalman filter (10.4) we require in addition the variables $\dot{Z}_t$ and $\dot{T}_t$ and they are given by

$$\dot{Z}_t = (\alpha_{2t}\,,\, \alpha_{1t}\,,\, 0)', \qquad \dot{T}_t = T_t.$$

Given these variables, we can carry out the computations for the extended Kalman filter (10.4) which needs to be initialised by

$$a_1 = 0, \qquad P_1 = \mathrm{diag}\left[\kappa\,,\, (1-\rho^2)^{-1}\sigma_\omega^2\,,\, (1-\rho^2)^{-1}\sigma_\omega^2\right],$$

where $\kappa \to \infty$. The initialisation of the extended Kalman filter needs to be modified since $\kappa \to \infty$ in P_1. The exact initialisation of the extended Kalman filter can be developed by following the treatment presented in Chapter 5 for the linear Gaussian case. We do not discuss these matters here further but refer to Koopman and Lee (2009) for a detailed derivation and discussion. An approximation to initialisation is to replace κ by a large value, say 10^7.

10.2.2 Power growth model

An alternative to the local linear trend model (3.2) discussed in Section 3.2 is the power growth model as given by

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big), \\ \mu_{t+1} &= \mu_t^{1+\nu_t} + \xi_t, & \xi_t &\sim \mathrm{N}\big(0, \sigma_\xi^2\big), \\ \nu_{t+1} &= \rho\nu_t + \zeta_t, & \zeta_t &\sim \mathrm{N}\big(0, \sigma_\zeta^2\big), \end{aligned} \tag{10.5}$$

where ρ is a discounting factor with $0 < \rho < 1$ for observation y_t and trend component μ_t with a stationary slope term ν_t. For this model we have a linear observation equation and a partly nonlinear state equation. The nonlinear state space model (10.1), with state vector $\alpha_t = (\mu_t\,,\, \nu_t)'$, has system matrices

$$Z_t(\alpha_t) = (1, 0), \qquad T_t(\alpha_t) = (\mu_t^{1+\nu_t}\,,\, \rho\nu_t)', \qquad Q_t(\alpha_t) = \mathrm{diag}(\sigma_\xi^2\,,\, \sigma_\zeta^2),$$

and with $H_t(\alpha_t) = \sigma_\varepsilon^2$ and $R_t(\alpha_t) = I_2$. The extended Kalman filter then relies only on $\dot{Z}_t$ and $\dot{T}_t$ which are given by

$$\dot{Z}_t = Z_t, \qquad \dot{T}_t = \begin{pmatrix} (\nu_t + 1)\mu_t^{\nu_t} & \mu_t^{1+\nu_t} \log \mu_t \\ 0 & \rho \end{pmatrix}.$$

The initial condition for the filter is given by

$$a_1 = 0, \qquad P_1 = \text{diag}\left[\kappa\, , \, (1 - \rho^2)^{-1}\sigma_\nu^2\right],$$

where $\kappa \to \infty$.

10.3 The unscented Kalman filter

The second approximate filter we consider for non-Gaussian nonlinear models is the *unscented Kalman filter* (UKF). It is based on a radically different idea from the linearisation used for the EKF. The idea can be most easily understood by considering its application to a simpler problem than the state space filtering problem. We therefore consider first in Subsection 10.3.1 the *unscented transformation* for vector functions of random vectors. The derivation of the UKF is given in Subsection 10.3.2. Further improvements of the basic unscented transformation are discussed in Subsection 10.3.3. The accuracy of the improved UKF in comparison with the EKF and standard UKF is investigated in Subsection 10.3.4.

10.3.1 The unscented transformation

Assume we have a $p \times 1$ random vector y which is a known nonlinear function

$$y = f(x), \tag{10.6}$$

of an $m \times 1$ random vector x with density $x \sim \text{N}(\bar{x}, P_{xx})$ and that we wish to find an approximation to the density of y. Julier and Uhlmann (1997) suggest that we proceed as follows: choose a set of *sigma points* denoted by $x_0, x_1, \ldots, x_{2m+1}$ with associated *sigma weights* denoted by $w_0, w_1, \ldots, w_{2m+1}$ where each $w_i > 0$ such that

$$\sum_{i=0}^{2m} w_i = 1, \qquad \sum_{i=0}^{2m} w_i x_i = \bar{x}, \qquad \sum_{i=0}^{2m} w_i (x_i - \bar{x})(x_i - \bar{x})' = P_{xx}. \tag{10.7}$$

In effect, we are approximating the continuous density $f(x)$ by a discrete density at points $x_0, x_1, \ldots, x_{2m+1}$ whose mean vector and variance matrix are the same as those of density $f(x)$. We then define $y_i = f(x_i)$, for $i = 0, \ldots, 2m$ and take

$$\bar{y} = \sum_{i=0}^{2m} w_i y_i, \qquad P_{yy} = \sum_{i=0}^{2m} w_i (y_i - \bar{y})(y_i - \bar{y})', \tag{10.8}$$

as our estimates of $\mathrm{E}(y)$ and $\mathrm{Var}(y)$, respectively. It is shown in Julier and Uhlmann (1997) that the lower order terms of the Taylor expansions of (10.6) can be made equal to the corresponding terms of the Taylor expansions of the true moments $\mathrm{E}(y)$ and $\mathrm{Var}(y)$ for any smooth function $f(\cdot)$. If more moments of x are known, for instance by assuming normality, it is possible to approximate more moments of y and to approximate the lower moments with greater precision.

There are many ways in which the sigma points and weights could be chosen subject to the constraints (10.7); Julier and Uhlmann (1997) suggest the very simple form

$$x_0 = \bar{x}, \qquad x_i = \bar{x} + \lambda\sqrt{P_{xx,i}}\,, \qquad x_{i+m} = \bar{x} - \lambda\sqrt{P_{xx,i}}\,,$$

with weights w_0 and

$$w_i = w_{i+m} = \frac{1-w_0}{2m}, \qquad i = 1, \ldots, m,$$

where λ is a scalar and $\sqrt{P_{xx,i}}$ is the ith column of a matrix square root of P_{xx} that can be obtained by, for example, the Cholesky decomposition of a symmetric matrix. The constant λ is determined by constraints as follows. The constraints

$$\sum_{i=0}^{2m} w_i = 1, \qquad \sum_{i=0}^{2m} w_i x_i = \bar{x},$$

are obviously satisfied. Substituting the x_i's and w_i's in the third item of (10.7) gives

$$\frac{1-w_0}{2m}\sum_{i=1}^{2m}\lambda^2\left(\sqrt{P_{xx,i}}\right)\left(\sqrt{P_{xx,i}}\right)' = P_{xx},$$

from which we deduce that

$$\lambda^2 = \frac{m}{1-w_0}.$$

By taking $w_0 = k \;/\; (m+k)$ for some value k, we obtain $\lambda^2 = m + k$ and

$$\begin{aligned} x_0 &= \bar{x}, & w_0 &= k \;/\; (m+k), \\ x_i &= \bar{x} + \sqrt{m+k}\sqrt{P_{xx,i}}, & w_i &= 1 \;/\; 2(m+k), \\ x_{i+m} &= \bar{x} - \sqrt{m+k}\sqrt{P_{xx,i}}, & w_{i+m} &= 1 \;/\; 2(m+k), \end{aligned} \tag{10.9}$$

for $i = 1, \ldots, m$. Julier and Uhlmann (1997) argue that 'k provides an extra degree of freedom to "fine tune" the higher order moments of the approximation, and can be used to reduce the overall prediction errors. When x is assumed Gaussian, a useful heuristic is to select $m + k = 3$.' While there are many possible alternatives we shall adopt this proposal for applications of the UKF to our state space model.

Julier and Uhlmann (1997) call this *the unscented transformation*. By examining Taylor expansions, they show that under appropriate conditions the unscented transformation estimates of the mean vector and the variance matrix of y are accurate to the second order of approximation.

10.3.2 Derivation of the unscented Kalman filter

We now construct a filter based on the unscented transformation. As for the EKF, we suppose that we have a series of observation vectors, $y_1, \dots, y_n$, with corresponding state vectors, $\alpha_1, \dots, \alpha_n$, which we assume are generated by the nonlinear non-Gaussian state space model (10.1). We carry out the construction in two stages, first the *updating stage* in which a new observation y_t has arrived and we wish to estimate $a_{t|t} = \mathrm{E}(\alpha_t|Y_t)$ and $P_{t|t} = \mathrm{Var}(\alpha_t|Y_t)$ given $a_t = \mathrm{E}(\alpha_t|Y_{t-1})$ and $P_t = \mathrm{Var}(\alpha_t|Y_{t-1})$, and secondly the *prediction stage* in which we wish to estimate a_{t+1} and P_{t+1} given $a_{t|t}$ and $P_{t|t}$, for $t = 1, \dots, n$. In order to keep the notation simple we shall use the same symbols for estimates and the quantities being estimated. We shall apply the unscented transformation separately in the two stages.

For the updating stage, let $\bar{y}_t = \mathrm{E}(y_t|Y_{t-1})$ and $v_t = y_t - \bar{y}_t$. The application of Lemma 1 in Section 4.2 is exact for the linear Gaussian state space model of Part I. In the case of non-Gaussian and nonlinear models Lemma 1 can still be applied but it provides only approximate relations. We use equation (4.2) of Lemma 1 in Section 4.2 to provide the approximate relation

$$\mathrm{E}(\alpha_t|Y_t) = \mathrm{E}(\alpha_t|Y_{t-1}) + P_{\alpha v,t} P_{vv,t}^{-1} v_t,$$

that is,

$$a_{t|t} = a_t + P_{\alpha v,t} P_{vv,t}^{-1} v_t, \tag{10.10}$$

where $P_{\alpha v,t} = \mathrm{Cov}(\alpha_t, v_t)$ and $P_{vv,t} = \mathrm{Var}(v_t)$ in the conditional joint distribution of α_t and v_t given Y_{t-1}, for $t = 1, \dots, n$. Similarly, we use equation (4.3) of Lemma 1 to provide the approximate relation

$$\mathrm{Var}(\alpha_t|Y_t) = \mathrm{Var}(\alpha_t|Y_{t-1}) - P_{\alpha v,t} P_{vv,t}^{-1} P'_{\alpha v,t},$$

that is,

$$P_{t|t} = P_t - P_{\alpha v,t} P_{vv,t}^{-1} P'_{\alpha v,t}. \tag{10.11}$$

Moreover, since the observation equation of model (10.1) is $y_t = Z(\alpha_t) + \varepsilon_t$, we have

$$\bar{y}_t = E\left[Z_t(\alpha_t)|Y_{t-1}\right]. \tag{10.12}$$

We proceed to estimate $\bar{y}_t$, $P_{\alpha v,t}$ and $P_{vv,t}$ by the unscented transformation. Define the sigma points and weights as follows,

$$\begin{aligned} x_{t,0} &= a_t, & w_0 &= k \;/\; (m+k), \\ x_{t,i} &= a_t + \sqrt{m+k} P^*_{t,i}, & w_i &= 1 \;/\; 2(m+k), \\ x_{t,i+m} &= a_t - \sqrt{m+k} P^*_{t,i}, & w_{i+m} &= 1 \;/\; 2(m+k), \end{aligned} \tag{10.13}$$

where m is the dimensionality of the state vector α_t and $P^*_{t,i}$ is the ith column of the square root matrix P^*_t obtained from a Cholesky decomposition $P_t = P^*_t P^{*\prime}_t$ for $i = 1, \dots, m$. We then take

$$\begin{aligned} \bar{y}_t &= \textstyle\sum_{i=0}^{2m} w_i Z_t(x_{t,i}), \\ P_{\alpha v,t} &= \textstyle\sum_{i=0}^{2m} w_i (x_{t,i} - a_t) \left[Z_t(x_{t,i}) - \bar{y}_t \right], \\ P_{vv,t} &= \textstyle\sum_{i=0}^{2m} w_i \left[Z_t(x_{t,i}) - \bar{y}_t \right] \left[Z_t(x_{t,i}) - \bar{y}_t \right]' + H_t(x_{t,i}), \end{aligned} \tag{10.14}$$

for $t = 1, \dots, n$. Taking $v_t = y_t - \bar{y}_t$, these are substituted in (10.10) and (10.11) to give $a_{t|t}$ and $P_{t|t}$, respectively.

To implement the prediction stage of the filter we first notice that the state equation of model (10.1) is $\alpha_{t+1} = T_t(\alpha_t) + R_t(\alpha_t)\eta_t$ where η_t is independent of α_t with vector mean zero and variance matrix $Q_t(\alpha_t)$. We then have

$$a_{t+1} = \mathrm{E}(\alpha_{t+1}|Y_t) = \mathrm{E}\left[T_t(\alpha_t)|Y_t) \right], \tag{10.15}$$

and

$$P_{t+1} = \mathrm{Var}(\alpha_{t+1}|Y_t) = \mathrm{Var}\left[T_t(\alpha_t)|Y_t) \right] + \mathrm{E}\left[R_t(\alpha_t) Q_t(\alpha_t) R_t(\alpha_t)'|Y_t) \right], \tag{10.16}$$

for $t = 1, \dots, n$. Define new $x_{t,0}, \dots, x_{t,2m}$ by relations (10.13) with a_t replaced by $a_{t|t}$ and P_t replaced by $P_{t|t}$. From these x_i's and the values for $w_0, \dots, w_{2m}$ from (10.13), we take

$$\begin{aligned} a_{t+1} &= \textstyle\sum_{i=0}^{2m} w_i T_t(x_{t,i}), \\ P_{t+1} &= \textstyle\sum_{i=0}^{2m} w_i \left[T_t(x_{t,i}) - a_{t+1} \right] \left[T_t(x_{t,i}) - a_{t+1} \right]' \\ &\quad + \textstyle\sum_{i=0}^{2m} w_i R_t(x_{t,i}) Q_t(x_{t,i}) R_t(x_{t,i})'. \end{aligned} \tag{10.17}$$

The filter defined by the relations (10.10), (10.11), (10.14) and (10.17) is called the *unscented Kalman filter* (UKF).

In terms of Taylor expansions, the unscented transform is accurate to the second order for the mean estimate. This result applies therefore to the estimate of the mean of the state vector by the UKF. The approximation of the mean by the EKF estimate is only accurate to the first order. However, the variance matrix is estimated to the second order of approximation by both the UKF and the EKF.

10.3.3 Further developments of the unscented transform

The original transformation of Julier and Uhlmann (1997) based on (10.9) is simple in the sense that we only take the two sigma points

$$\bar{x} \pm \sqrt{m+k}\sqrt{P_{xx,i}},$$

for each ith element of the vector x. We now consider whether we can improve efficiency by increasing the number of sigma points for each i and also by allocating relatively higher weight to sigma points that correspond to higher densities. For this purpose we suggest taking $2\,q$ sigma points for each i by considering

$$\begin{aligned} x_{ij} &= \bar{x} + \lambda\xi_j\sqrt{P_{xx,i}} \qquad i = 1,\ldots,m, \\ x_{ij} &= \bar{x} - \lambda\xi_j\sqrt{P_{xx,i}} \qquad i = m+1,\ldots,2m, \end{aligned} \qquad j = 1,\ldots,q, \tag{10.18}$$

together with $x_0 = \bar{x}$. For example, for $q = 2$, we could take $\xi_1 = 1$ and $\xi_2 = 2$ or $\xi_1 = \frac{1}{2}$ and $\xi_2 = \frac{3}{2}$, and for $q = 4$, we could take $\xi_1 = \frac{1}{2}$, $\xi_2 = 2$, $\xi_3 = \frac{3}{2}$ and $\xi_4 = 2$. Associated with these sigma points, we take the weights

$$w_{ij} = w_{m+i,j} = w\varphi(\xi_j), \qquad i = 1,\ldots,m, \qquad j = 1,\ldots,q,$$

together with w_0, where w and w_0 are to be determined and where $\varphi(\cdot)$ is the standard normal density. The constraint that the weights must sum to one,

$$w_0 + 2\sum_{i=1}^{m}\sum_{j=1}^{q} w_{ij} = 1,$$

leads to

$$w_0 + 2m\,w\sum_{j=1}^{q}\varphi(\xi_j) = 1,$$

which gives

$$w = \frac{1 - w_0}{2m\sum_{j=1}^{q}\varphi(\xi_j)}. \tag{10.19}$$

It also follows that

$$\sum_{i=0}^{2m}\sum_{j=1}^{q} w_{ij}x_{ij} = \left[w_0 + 2m\,w\sum_{j=1}^{q}\varphi(\xi_j)\right]\bar{x} = \bar{x}.$$

The constraint

$$2\sum_{i=1}^{m}\sum_{j=1}^{q} w_{ij}(x_{ij}-\bar{x})(x_{ij}-\bar{x})' = 2\sum_{i=1}^{m}\sum_{j=1}^{q} w_{ij}\lambda^2\xi_j^2 P_{xx,i}^{*} P_{xx,i}^{*\,\prime} = P_{xx},$$

where

$$w_{ij} = \frac{(1 - w_0)\varphi(\xi_j)}{2m\sum_{l=1}^{q}\varphi(\xi_l)}$$

gives

$$\frac{2(1 - w_0)\lambda^2}{2m\sum_{l=1}^{q}\varphi(\xi_l)} m\sum_{j=1}^{q}\varphi(\xi_j)\xi_j^2 = 1, \tag{10.20}$$

and hence

$$\lambda^2 = \frac{\sum_{j=1}^q \varphi(\xi_j)}{(1-w_0)\sum_{j=1}^q \varphi(\xi_j)\xi_j^2}. \tag{10.21}$$

We now seek a further constraint by equalising fourth moments in order to provide a value for w_0. Denote the scalar elements of the vector x by x_ℓ^* for $\ell = 1, \ldots, m$ so that $x = (x_1^*, \ldots, x_m^*)'$. Let $\bar{x}_\ell^* = \mathrm{E}(x_\ell^*)$ for $\ell = 1, \ldots, m$. For simplicity, take the case

$$P_{xx} = \mathrm{Var}(x) = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right),$$

and assume that x is normally distributed. Thus $x_\ell^* \sim \mathrm{N}(\bar{x}_\ell^*, \sigma_\ell^2)$ for $\ell = 1, \ldots, m$. Denote the ℓth element of x_{ij} by $x_{ij\ell}$ and assign to it the weight w_{ij}. From (10.18) we have

$$\begin{aligned} x_{ij\ell} &= \bar{x}_\ell^* + \lambda\xi_j\sigma_\ell, \qquad i = 1, \ldots, m, \\ &= \bar{x}_\ell^* - \lambda\xi_j\sigma_\ell, \qquad i = m+1, \ldots, 2m, \end{aligned} \tag{10.22}$$

for $j = 1, \ldots, q$ and $\ell = 1, \ldots, m$. Let us impose the fourth moment constraint

$$\begin{aligned} \mathrm{E}\left[\sum_{\ell=1}^m (x_\ell^* - \bar{x}_\ell^*)^4\right] &= 2\sum_{\ell=1}^m\sum_{i=1}^m\sum_{j=1}^q w_{ij}\mathrm{E}\left[(x_\ell^* - \bar{x}_\ell^*)^4\right] \\ &= 2\sum_{\ell=1}^m\sum_{i=1}^m\sum_{j=1}^q w\varphi(\xi_j)\left(\lambda\xi_j\sigma_\ell\right)^4 \\ &= 2m\, w\lambda^4 \sum_{\ell=1}^m\sum_{j=1}^q \varphi(\xi_j)\xi_j^4\sigma_\ell^4. \end{aligned}$$

Since the fourth moment of $\mathrm{N}(0, \sigma_\ell^2)$ is $3\sigma_\ell^4$, and using (10.19) and (10.21), we have

$$3\sum_{\ell=1}^m \sigma_\ell^4 = \frac{1-w_0}{\sum_{j=1}^q \varphi(\xi_j)} \frac{\left[\sum_{j=1}^q \varphi(\xi_j)\right]^2 \sum_{j=1}^q \varphi(\xi_j)\xi_j^4 \sum_{\ell=1}^m \sigma_\ell^4}{(1-w_0)^2[\sum_{j=1}^q \varphi(\xi_j)\xi_j^2]^2}, \tag{10.23}$$

giving

$$3 = \frac{\sum_{j=1}^q \varphi(\xi_j)\sum_{j=1}^q \varphi(\xi_j)\xi_j^4}{(1-w_0)\left[\sum_{j=1}^q \varphi(\xi_j)\xi_j^2\right]^2}, \tag{10.24}$$

and hence

$$w_0 = 1 - \frac{\sum_{j=1}^q \varphi(\xi_j)\sum_{j=1}^q \varphi(\xi_j)\xi_j^4}{3\left[\sum_{j=1}^q \varphi(\xi_j)\xi_j^2\right]^2}. \tag{10.25}$$

We use this value as an approximation when P_{xx} is not diagonal.

We have thus developed an expanded set of sigma points spread over a wider area than (10.9) over the support of the distribution, while maintaining the correspondence between moments of the distribution over the sigma points and the population moments.

10.3.4 Comparisons between EKF and UKF

In this section we provide a comparison of the estimation performance of three nonlinear filtering approaches. We generate 10,000 replications from the (partly) nonlinear state space model (10.1) with a stationary autoregressive process for the univariate state α_t, that is

$$y_t = Z_t(\alpha_t) + \varepsilon_t, \qquad \alpha_{t+1} = 0.95\alpha_t + \eta_t,$$

with error terms $\varepsilon_t \sim \mathrm{N}(0, 0.01)$ and $\eta_t \sim \mathrm{N}(0, 0.01)$ for $t = 1, \ldots, 100$. We consider different nonlinear transformations for $Z_t(\alpha_t)$. Table 10.1 shows the mean of the sum of squared prediction errors for different choices of the $Z(\alpha_t)$ function and for the filters EKF, UKF and the modified UKF. The results show that the unscented filters generally provide more accurate state predictions than the extended filter. When we increase the order of the polynomial transformation, the accuracies diverge: at lower orders the two unscented filters have a similar performance, while at high orders the modified UKF performs significantly better. We achieve similar improvements with the $\exp(\alpha_t)$ and $\sin(\alpha_t) + 1.1$ transformations although they are small and comparable with transformations such as α_t^2 and α_t^3. The $\log(\alpha_t + 6)$ transformation can be appropriately handled by the EKF since the UKF does not provide much improvement.

Table 10.1 Mean squared prediction errors for three nonlinear filtering methods.

$Z(\alpha_t)$	EKF	UKF	MUKF
α_t^2	55.584	30.440	30.440
α_t^3	16.383	10.691	10.496
α_t^4	46.048	43.484	28.179
α_t^5	107.23	17.019	13.457
α_t^6	147.62	92.216	30.651
α_t^7	1468.1	85.311	24.175
α_t^8	1641.1	347.84	39.996
$\exp(\alpha_t)$	13.177	11.917	11.866
$\sin(\alpha_t) + 1.1$	32.530	24.351	23.514
$\log(\alpha_t + 6)$	38.743	38.695	38.695

The EKF refers to the extended Kalman filter of Section 10.2, UKF refers to the unscented Kalman filter of Subsection 10.3.1 and MUKF refers to the modified UKF method of Subsection 10.3.3 with $q = 4$.

10.4 Nonlinear smoothing

Smoothing is extensively discussed in Chapter 4 for the linear Gaussian model. Approximate filtering methods for nonlinear models can be based on the extended Kalman filter and on the unscented Kalman filter. Since both approximations are associated with the Kalman filter recursions, it can be anticipated that methods of smoothing for the linear Gaussian model can be adapted similarly. We discuss some details for both approaches.

10.4.1 Extended smoothing

The linearisation technique that we employed in Section 10.2 to provide the extended Kalman filter can be used to obtain an approximate state smoother. We simply take the linearised form (10.3) of the nonlinear model (10.1) and apply the EKF (10.4). The system matrices of the linearised state space form vary over time and depend on predicted and filtered estimates of the state vector at time t. The backward smoothing methods as described in Section 4.4 can be used as approximations to smoothing. More specifically, the recursions for the extended state smoother are given by

$$r_{t-1} = \dot{Z}_t' F_t^{-1} v_t + L_t' r_t, \qquad \hat{\alpha}_t = a_t + P_t r_{t-1}, \qquad t = n, \ldots, 1, \quad (10.26)$$

where F_t, v_t,, L_t and P_t are obtained from the extended Kalman filter (10.4) while the recursion is initiated with $r_n = 0$.

For the sole purpose of smoothing, the linearised form (10.3) can be considered once more, in a second round of Kalman filtering and smoothing. We then evaluate $\dot{T}_t$ and $\dot{Z}_t$ in (10.2) once more but now at $\alpha_t = \hat{\alpha}_t$, for $t = 1, \ldots, n$. The Kalman filter and smoother equations are used to obtain the new estimates $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$. We expect that the linear approximation based on $\hat{\alpha}_t$ is more accurate than the approximation based on a_t and $a_{t|t}$.

10.4.2 Unscented smoothing

The unscented filtering method can be modified in different ways to obtain an unscented smoothing method. For example, the two filter formula method of Subsection 4.6.4 can be advocated. The UKF is applied to the time series once and the UKF is also applied to the same time series in reverse order. The UKF estimates are then combined as in (4.74) to obtain the smoothed estimates. This approach only leads to an approximation of $\mathrm{E}(\alpha_t|Y_n)$ for a nonlinear state space model.

An alternative and our preferred approximation to a smoothing algorithm based on the unscented transformation can be derived from Lemma 1 of Chapter 4. Its development is similar to the classical fixed interval smoother of Section 4.4. Assume that the α_t and α_{t+1} conditional on Y_t are jointly normally distributed with

$$\mathrm{E}(\alpha_t|Y_t) = a_{t|t}, \qquad \mathrm{E}(\alpha_{t+1}|Y_t) = a_{t+1}, \qquad (10.27)$$

$$\mathrm{Var}(\alpha_t|Y_t) = P_{t|t}, \qquad \mathrm{Var}(\alpha_{t+1}|Y_t) = P_{t+1}, \tag{10.28}$$

and

$$\mathrm{Cov}(\alpha_t, \alpha_{t+1}|Y_t) = C_{t+1}. \tag{10.29}$$

From Lemma 1, we obtain the approximation

$$\begin{aligned} \hat{\alpha_t} &= \mathrm{E}(\alpha_t|Y_n) \\ &= \mathrm{E}(\alpha_t|\alpha_{t+1}, Y_t) \\ &= a_{t|t} + C_{t+1}P_{t+1}^{-1}\big(\hat{\alpha}_{t+1} - a_{t+1}\big), \end{aligned} \tag{10.30}$$

for $t = 1, \ldots, n$, where a_{t+1}, $a_{t|t}$ and P_{t+1} have been computed by the UKF, and

$$C_{t+1} = \sum_{i=0}^{2m} w_i\big(x_{t,i} - a_t\big)\big[T_t(x_{t,i}) - a_{t+1}\big]'. \tag{10.31}$$

These results imply a backward recursive algorithm for smoothing. Using the same arguments, we can compute the approximated smoothed state variance. A recent discussion of unscented smoothing is provided by Särkkä and Hartikainen (2010).

10.5 Approximation via data transformation

Nonlinear and non-Gaussian models can sometimes be successfully approximated by transforming observations into forms for which a linear Gaussian model can be employed as an approximation. This approach of approximation via data transformation is usually part of an *ad hoc* solution. In case of the nonlinear model (9.36), that is

$$y_t = Z_t(\alpha_t) + \varepsilon_t, \qquad \alpha_{t+1} = T_t\alpha_t + R_t\eta_t,$$

for $t = 1, \ldots, n$, we can consider a function $Z_t^{-}(\alpha_t)$ such that

$$Z_t^{-}\left[Z_t(\alpha_t)\right] = c_t^a + Z_t^a\alpha_t,$$

where vectors c_t^a and matrix Z_t^a have appropriate dimensions. The approximation is then achieved by applying the transformation $Z_t^{-}(\alpha_t)$ since

$$Z_t^{-}\left[Z_t(\alpha_t) + \varepsilon_t\right] \approx c_t^a + Z_t^a\alpha_t + u_t,$$

where u_t is an error term. A typical univariate example is to take $Z_t(\alpha_t) = \exp(\alpha_t)$ such that $Z_t^{-}(\alpha_t) = \log(\alpha_t)$.

10.5.1 Partly multiplicative decompositions

Consider the multiplicative model for the univariate observation y_t given by

$$y_t = \mu_t \times \gamma_t + \varepsilon_t,$$

for $t = 1, \dots, n$, where μ_t is an unobserved trend component and γ_t is an unobserved seasonal or cycle component; see Section 3.2. By taking logs of y_t we obtain approximately

$$\log(y_t) \approx \log(\mu_t) + \log(\gamma_t) + u_t,$$

where appropriate dynamic properties are given to the components $\log(\mu_t)$ and $\log(\gamma_t)$.

10.5.2 Stochastic volatility model

Consider the basic SV model (9.26) of Section 9.5 as given by

$$y_t = \mu + \sigma \exp\left(\frac{1}{2}\theta_t\right)\varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0,1),$$

for $t = 1, \dots, n$, where the log-volatility is modelled as an autoregressive process (9.27), that is

$$\theta_{t+1} = \phi\theta_t + \eta_t, \qquad \eta_t \sim \mathrm{N}\big(0, \sigma_\eta^2\big)$$

and where the disturbances ε_t and η_t are mutually and serially uncorrelated. To obtain an approximate solution based on a linear model, we can transform the observations y_t as follows

$$\log y_t^2 = \kappa + \theta_t + \xi_t, \qquad t = 1, \dots, n, \tag{10.32}$$

where

$$\kappa = \log\sigma^2 + \mathrm{E}\big(\log\varepsilon_t^2\big), \qquad \xi_t = \log\varepsilon_t^2 - \mathrm{E}\big(\log\varepsilon_t^2\big). \tag{10.33}$$

The noise term ξ_t is not normally distributed but the model for $\log y_t^2$ is linear and therefore we can proceed approximately with the linear techniques of Part I. This approach is taken by Harvey, Ruiz and Shephard (1994) who call the procedure for parameter estimation based on it *quasi-maximum likelihood* (QML). Parameter estimation is done via the Kalman filter; smoothed estimates of the volatility component, θ_t, can be constructed and forecasts of volatility can be generated. One of the attractions of the QML approach is that it can be carried out straightforwardly using the readily available software program *STAMP*; see Koopman, Harvey, Doornik and Shephard (2010).

10.6 Approximation via mode estimation

The treatments for the extended and unscented filters are principally motivated by nonlinear models of the form (10.1) where the state vector is the argument of the nonlinear functions in the model. A general class of models can be formulated by

$$y_t \sim p(y_t|\alpha_t), \qquad \alpha_{t+1} = T_t(\alpha_t) + R_t(\alpha_t)\eta_t, \tag{10.34}$$

where $p(y_t|\alpha_t)$ is the observation density conditional on the state vector α_t which can evolve over time depending on nonlinear functions of the state vector but which in practice is often linear and Gaussian. In case the observation density has a possibly nonlinear mean function $Z_t(\alpha_t)$ and a possibly nonlinear variance function $H_t(\alpha_t)$, the models (10.1) and (10.34) are equivalent. The focus of model (10.34) is, however, on the non-Gaussian feature of $p(y_t|\alpha_t)$ rather than on the nonlinear functions of the model. The extended and unscented filters do not account for the non-Gaussian properties of the model other than their first two moments: the mean and variance functions. We will show that computing the mode of the state vector conditional on all observations will lead to an approximating linear Gaussian state space model.

We first discuss the case where $p(y_t|\alpha_t)$ can be any density but where α_t evolves linearly over time with a Gaussian error vector η_t. We further assume that the signal vector θ_t is a linear function of the state vector and that it is sufficient for the observation density; see Section 9.2 for a discussion on this class of models. In other words, density $p(y_t|\alpha_t)$ is equivalent to $p(y_t|\theta_t)$. We obtain the non-Gaussian model specification

$$y_t \sim p(y_t|\theta_t), \quad \theta_t = Z_t\alpha_t, \quad \alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \quad \eta_t \sim \mathrm{N}(0, Q_t), \tag{10.35}$$

for $t = 1, \ldots, n$. The computation of the mode of θ_t conditional on all observations will lead to an approximating linear Gaussian state space model. The standard Kalman filtering and smoothing methods can be used for analysis.

10.6.1 Mode estimation for the linear Gaussian model

In Section 4.13 we have shown that the linear Gaussian state space model can be expressed in matrix form. Since we wish to focus on the signal vector θ, where $\theta = (\theta_1', \ldots, \theta_n')'$ with signal $\theta_t = Z_t\alpha_t$ as defined in (9.5) for $t = 1, \ldots, n$, we write the observation equation in matrix form as

$$Y_n = \theta + \varepsilon, \qquad \theta = Z\alpha, \qquad \varepsilon \sim \mathrm{N}(0, H), \tag{10.36}$$

where $\alpha = T(\alpha_1^* + R\eta)$ and matrix H is block-diagonal. All vector and matrix definitions are given in Subsection 4.13.1, including definitions for matrices Z, H, T, R and Q in the equations (4.94)–(4.99) and including equations for the mean vector and variance matrix of θ, that is

$$\mathrm{E}(\theta) = \mu = ZTa_1^*, \qquad \mathrm{Var}(\theta) = \Psi = ZT(P_1^* + RQR')T'Z',$$

where a_1^* and P_1^* are defined below (4.100). It follows that the observation equation (10.36) can also be expressed as

$$Y_n = \mu + u, \qquad u \sim \mathrm{N}(0, \Sigma), \qquad \Sigma = \Psi + H.$$

Since the mode of Gaussian densities is equal to the mean it follows from Subsection 4.13.5 that the mode of the signal can be expressed by

$$\hat{\theta} = (\Psi^{-1} + H^{-1})^{-1}(\Psi^{-1}\mu + H^{-1}Y_n), \tag{10.37}$$

where the mode $\hat{\theta}$ is the value of θ that maximises the smoothed density $p(\theta|Y_n)$ for the linear Gaussian state space model.

10.6.2 Mode estimation for model with linear Gaussian signal

Here we aim to estimate the mode for the smoothed density of the class of models represented by (10.34) where the signal is linear and Gaussian; see Section 9.2. This class of models in matrix form is

$$Y_n \sim p(Y_n|\theta), \qquad \theta \sim \mathrm{N}(\mu, \Psi),$$

where

$$p(Y_n|\theta) = \prod_{t=1}^{n} p(y_t|\theta_t) = \prod_{t=1}^{n} p(\varepsilon_t).$$

The smoothed density $p(\theta|Y_n)$ does not have an explicit expression from which we can obtain the mode analytically. Therefore, we express the smoothed logdensity by

$$\log p(\theta|Y_n) = \log p(Y_n|\theta) + \log p(\theta) - \log p(Y_n), \tag{10.38}$$

and maximise this expression with respect to θ numerically using the Newton–Raphson method; see Nocedal and Wright (1999) for a general discussion of the Newton–Raphson method. The components of the smoothed density in (10.38) dependent on θ are the observation density $p(Y_n|\theta)$ and the signal density $p(\theta)$ as given by (4.111), that is

$$\log p(\theta) = \mathrm{N}(\mu, \Psi) = \text{constant} - \frac{1}{2}\log|\Psi| - \frac{1}{2}(\theta - \mu)'\Psi^{-1}(\theta - \mu), \tag{10.39}$$

since θ is linear and Gaussian. The density $p(Y_n)$ does not depend on θ.

For a given guess of the mode, say $\tilde{\theta}$, a new guess of the mode, say $\tilde{\theta}^+$, is obtained by solving a second-order Taylor expansion of $\log p(\theta|Y_n)$ around $\theta = \tilde{\theta}$. We have

$$\tilde{\theta}^+ = \tilde{\theta} - \left[\ddot{p}(\theta|Y_n)|_{\theta=\tilde{\theta}}\right]^{-1} \dot{p}(\theta|Y_n)|_{\theta=\tilde{\theta}}\,, \tag{10.40}$$

where

$$\dot{p}(\cdot|\cdot) = \frac{\partial \log p(\cdot|\cdot)}{\partial\theta}, \qquad \ddot{p}(\cdot|\cdot) = \frac{\partial^2 \log p(\cdot|\cdot)}{\partial\theta\partial\theta'}. \tag{10.41}$$

Given these definitions and the expressions (10.38) and (10.39) for $\log p(\theta|Y_n)$ and $\log p(\theta)$, respectively, we obtain

$$\dot{p}(\theta|Y_n) = \dot{p}(Y_n|\theta) - \Psi^{-1}(\theta - \mu), \qquad \ddot{p}(\theta|Y_n) = \ddot{p}(Y_n|\theta) - \Psi^{-1}. \tag{10.42}$$

The independence assumption (9.2) implies that

$$\log p(Y_n|\theta) = \sum_{t=1}^{n} \log p(y_t|\theta_t),$$

so that matrix $\ddot{p}(Y_n|\theta)$ is block-diagonal. More specifically, we have

$$\begin{aligned} \dot{p}(Y_n|\theta) &= \left[\dot{p}_1(y_1|\theta_1), \dots, \dot{p}_n(y_n|\theta_n)\right]', \\ \ddot{p}(Y_n|\theta) &= \mathrm{diag}\left[\ddot{p}_1(y_1|\theta_1), \dots, \ddot{p}_n(y_n|\theta_n)\right], \end{aligned} \tag{10.43}$$

where

$$\dot{p}_t(\cdot|\cdot) = \frac{\partial \log p(\cdot|\cdot)}{\partial \theta_t}, \qquad \ddot{p}_t(\cdot|\cdot) = \frac{\partial^2 \log p(\cdot|\cdot)}{\partial \theta_t \partial \theta_t'},$$

for $t = 1, \dots, n$.

By substitution of (10.42) into (10.40), the Newton–Raphson updating step (10.40) becomes

$$\begin{aligned} \tilde{\theta}^{+} &= \tilde{\theta} - \left\{ \ddot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}} - \Psi^{-1} \right\}^{-1} \left\{ \dot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}} - \Psi^{-1}(\tilde{\theta} - \mu) \right\} \\ &= \left(\Psi^{-1} + A^{-1} \right)^{-1} \left(A^{-1}x + \Psi^{-1}\mu \right), \end{aligned} \tag{10.44}$$

where

$$A = -\left\{ \ddot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}} \right\}^{-1}, \qquad x = \tilde{\theta} + A\, \dot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}}. \tag{10.45}$$

We note the similarity between (10.44) and (10.37). In case $\ddot{p}(Y_n|\theta)$ is negative definite for all θ, it is due the block-diagonality of matrix A as implied by (10.43) that the Kalman filter and smoother of Chapter 4 can be used to compute the next guess of the mode $\tilde{\theta}^{+}$, for a given current guess of $\tilde{\theta}$. This Kalman filter and smoother is based on the Gaussian state space model (10.36) with $Y_n = x$ and $H = A$. When the new estimate $\tilde{\theta}^{+}$ is computed, we can treat it as a new guess $\tilde{\theta} = \tilde{\theta}^{+}$ for which yet another new guess can be computed. This process can be repeated and constitutes the Newton–Raphson method for this application. Convergence is usually fast and only around ten iterations or less are needed in many cases of interest. After convergence, we have obtained the mode $\hat{\theta}$ with Hessian matrix $G = \ddot{p}(\theta|y)|_{\theta=\hat{\theta}} = -\Psi^{-1} - A^{-1}$, where A is evaluated at $\theta = \hat{\theta}$. It is shown that for an appropriately designed linear Gaussian model, the Kalman filter and smoother is able to compute the mode for the nonlinear non-Gaussian model (10.34).

The iterative approach for computing the mode was proposed by Shephard and Pitt (1997), Durbin and Koopman (1997) and So (2003, §2). The method is clearly not valid when $\ddot{p}(y|\theta)$ is not negative definite since this will imply that the variance matrix H of the linear Gaussian model (10.36) is non-negative definite. In other words, density $p(y|\theta)$ must be logconcave in θ. In the case $p(y|\theta)$ is not logconcave, the method of this section can still be adopted but the derivation must be based on other arguments; see Jungbacker and Koopman (2007).

When the mode of the state vector α_t is required for (10.34), we can repeat the above derivation with $p(\theta_t|Y_n)$ replaced by $p(\alpha_t|Y_n)$ and θ_t replaced by α_t. However, it is argued in Subsection 4.5.3 that state smoothing is computationally more involved than signal smoothing. Since we need to apply the Kalman filter and smoothing algorithms repeatedly as part of the Newton–Raphson method for computing the mode, the computation time is higher for obtaining the mode of the state compared to obtaining the mode of the signal. In this respect the result of Subsection 4.13.5 is relevant. It shows that once $\hat{\theta}$ is obtained, the estimate $\hat{\alpha}$ can be computed based on a single Kalman filter and smoother applied to model (4.114). It implies that once the signal mode $\hat{\theta}$ is computed (mean and mode are the same in a linear Gaussian model), it is not necessary to repeat a Newton–Raphson method to compute the mode for the state vector. We can simply formulate the linear Gaussian state space model (4.114) where the observations are replaced by $\hat{\theta}_t$ and where the observation noise is set to zero. The Kalman filter and smoothing recursions for the state vector compute the mode $\hat{\alpha}_t$ for $t = 1, \dots, n$.

10.6.3 Mode estimation by linearisation

The computation of the mode as described in Subsection 10.6.2 can alternatively be derived by matching the first and second derivatives of the smoothing densities $p(\theta|Y_n)$ and $g(\theta|Y_n)$ where $g(\theta|Y_n)$ refers to the smoothed approximating density of the linear Gaussian model. The logdensity $\log g(\theta|Y_n)$ can be decomposed as in (10.38), that is

$$\log g(\theta|Y_n) = \log g(Y_n|\theta) + \log g(\theta) - \log g(Y_n),$$

where $\log g(Y_n|\theta)$ is the logdensity of the observation equation and is defined by (4.105). We therefore obtain

$$\dot{g}(Y_n|\theta) = H^{-1}(Y_n - \theta), \qquad \ddot{g}(Y_n|\theta) = -H^{-1}, \tag{10.46}$$

where

$$\dot{g}(\cdot|\cdot) = \frac{\partial \log g(\cdot|\cdot)}{\partial \theta}, \qquad \ddot{g}(\cdot|\cdot) = \frac{\partial^2 \log g(\cdot|\cdot)}{\partial \theta \partial \theta'}. \tag{10.47}$$

Furthermore, since we assume a linear Gaussian signal θ, we have $g(\theta) = p(\theta)$ and is given by (4.111). Finally, $\log g(Y_n)$ is the logdensity of the observations and does not depend on θ. Matching the first two derivatives of the densities

$p(\theta|Y_n)$ and $g(\theta|Y_n)$, with respect to θ, is therefore equivalent to matching the model densities $g(Y_n|\theta)$ and $p(Y_n|\theta)$, that is

$$H^{-1}(Y_n - \theta) = \dot{p}(Y_n|\theta), \qquad -H^{-1} = \ddot{p}(Y_n|\theta). \tag{10.48}$$

Since the derivatives are functions of θ themselves, we solve the equations in (10.48) by iteration. For a given value of $\theta = \tilde{\theta}$, we equalise the derivatives in (10.48) by transforming the observation vector Y_n into x and the observation disturbance variance matrix H into A for the linear Gaussian model as

$$A = -\left\{ \ddot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}} \right\}^{-1}, \qquad x = \tilde{\theta} + A\, \dot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}}\,;$$

compare the definitions of x and A in (10.45). The application of the Kalman filter and smoother to the linear Gaussian model with observation vector $Y_n = x$ and variance matrix $H = A$, produces a new estimate $\tilde{\theta}^+$ of θ. We can replace $\tilde{\theta}$ by $\tilde{\theta}^+$ to linearise again in the form (10.48). This process leads to an iteration from which the final linearised model is the linear Gaussian model with the same conditional mode of θ given Y_n as the non-Gaussian nonlinear model.

We have shown that matching the first and second derivatives of the smoothed densities is equivalent to maximising the smoothed density of the signal $p(\theta|Y_n)$ with respect to θ. The value of θ at its maximum is the mode of θ. In the next section we will approximate more general nonlinear non-Gaussian state space models by matching the first and second derivatives of the smoothed densities with respect to the state vector α.

In case the observations can be represented by the signal plus noise model of the form (9.7) in Section 9.2, linearisation based on matching the first derivatives only may also be appropriate. We have $y_t = \theta_t + \varepsilon_t$ with $\varepsilon_t \sim p(\varepsilon_t)$ such that $p(y_t|\theta_t) = p(\varepsilon_t)$. We shall assume that y_t is univariate because it is an important case in practice and it simplifies the treatment. By matching the first derivative between $g(Y_n|\theta)$ and $p(Y_n|\theta)$ only, we obtain

$$H^{-1}(Y_n - \theta) = \dot{p}(\varepsilon), \tag{10.49}$$

where $\dot{p}(\varepsilon) = \dot{p}(Y_n|\theta)$ as defined in (10.41). For a given value of $\theta = \tilde{\theta}$ and $\varepsilon = \tilde{\varepsilon}$ with $\tilde{\varepsilon} = Y_n - \tilde{\theta}$, we equalise the first derivatives in (10.49) and transform the observation disturbance variance matrix H into A for the linear Gaussian signal plus noise model, to obtain

$$A_t = (y_t - \tilde{\theta}_t)\dot{p}(\varepsilon_t)^{-1} \bigg|_{\varepsilon_t = y_t - \tilde{\theta}_t}\,,$$

where A_t is the tth diagonal element of A. The observation does not need transformation so that $x_t = y_t$ for $t = 1, \ldots, n$.

10.6.4 Mode estimation for exponential family models

An important application of these results is to observations from the exponential family of distributions. For this class of models we can compute the mode as described in Subsection 10.6.2. For density (9.3) we have

$$\log p(y_t|\theta_t) = y_t'\theta_t - b_t(\theta_t) + c_t(y_t). \tag{10.50}$$

For a given value $\tilde{\theta} = (\tilde{\theta}_1', \ldots, \tilde{\theta}_n')'$ for θ, the resulting derivatives are given by

$$\dot{p}(y_t|\theta_t) = y_t - \dot{b}_t, \qquad \ddot{p}(y_t|\theta_t) = -\ddot{b}_t,$$

where

$$\dot{b}_t = \left.\frac{\partial b_t(\theta_t)}{\partial \theta_t}\right|\theta_t = \tilde{\theta}_t\ , \qquad \ddot{b}_t = \left.\frac{\partial^2 b_t(\theta_t)}{\partial \theta_t \partial \theta_t'}\right|\theta_t = \tilde{\theta}_t\ ,$$

for $t = 1, \ldots, n$. These values can be substituted in (10.45) to obtain a solution for the case where signal θ is linear and Gaussian, that is

$$A_t = \ddot{b}_t^{-1}, \qquad x_t = \tilde{\theta}_t + \ddot{b}_t^{-1} y_t - \ddot{b}_t^{-1}\dot{b}_t,$$

where A_t is the tth element of the diagonal matrix A and x_t is the tth element of vector x; (block-)diagonal matrix A and vector x are defined in (10.45). Since, as shown in Section 9.3, $\ddot{b}_t = \mathrm{Var}(y_t|\theta_t)$, it is positive definite in nondegenerate cases, so for the exponential family, the method of computing the mode can always be used.

As an example, for the Poisson distribution with density (9.12), we have

$$\log p(y_t|\theta_t) = y_t\theta_t - \exp\theta_t - \log y_t!,$$

so that $b_t(\tilde{\theta}_t) = \dot{b}_t = \ddot{b}_t = \exp(\tilde{\theta}_t)$. For computing the mode, we therefore take

$$A_t = \exp(-\tilde{\theta}_t), \qquad x_t = \tilde{\theta}_t + \exp(-\tilde{\theta}_t)y_t - 1,$$

for $t = 1, \ldots, n$. Other examples of expressions for A_t and x_t for a range of exponential family models are given in Table 10.2.

10.6.5 Mode estimation for stochastic volatility model

The mode estimation of the volatility signal in a stochastic volatility model should be based on the first two derivatives. For the basic SV model (9.26) we have

$$\log p(y_t|\theta_t) = -\frac{1}{2}\left[\log 2\pi\sigma^2 + \theta_t + z_t^2\exp(-\theta_t)\right],$$

where $z_t = (y_t - \mu)/\sigma$. It follows that

$$\dot{p}_t = -\frac{1}{2}\left[1 - z_t^2\exp(-\theta_t)\right], \qquad \ddot{p}_t = -\frac{1}{2}z_t^2\exp(-\theta_t).$$

Table 10.2 Approximating model details for exponential family models.

Distribution		
Poisson	b_t	$\exp\theta_t$
	$\dot{b}_t$	$\exp\theta_t$
	$\ddot{b}_t$	$\exp\theta_t$
	$\ddot{b}_t^{-1}\dot{b}_t$	1
binary	b_t	$\log(1+\exp\theta_t)$
	$\dot{b}_t$	$\exp\theta_t(1+\exp\theta_t)^{-1}$
	$\ddot{b}_t$	$\exp\theta_t(1+\exp\theta_t)^{-2}$
	$\ddot{b}_t^{-1}\dot{b}_t$	$1+\exp\theta_t$
binomial	b_t	$k_t\log(1+\exp\theta_t)$
	$\dot{b}_t$	$k_t\exp\theta_t(1+\exp\theta_t)^{-1}$
	$\ddot{b}_t$	$k_t\exp\theta_t(1+\exp\theta_t)^{-2}$
	$\ddot{b}_t^{-1}\dot{b}_t$	$1+\exp\theta_t$
negative binomial	b_t	$k_t\{\theta_t-\log(1-\exp\theta_t)\}$
	$\dot{b}_t$	$k_t(1-\exp\theta_t)^{-1}$
	$\ddot{b}_t$	$k_t\exp\theta_t(1-\exp\theta_t)^{-2}$
	$\ddot{b}_t^{-1}\dot{b}_t$	$\exp(-\theta_t)-1$
exponential	b_t	$-\log\theta_t$
	$\dot{b}_t$	$-\theta_t^{-1}$
	$\ddot{b}_t$	θ_t^{-2}
	$\ddot{b}_t^{-1}\dot{b}_t$	$-\theta_t$

The key variables x_t and A_t for the approximating model $x_t = \theta_t + u_t$ with $u_t \sim \mathrm{N}(0, A_t)$ are given by $x_t = \tilde{\theta}_t + \ddot{b}_t^{-1}y_t - \ddot{b}_t^{-1}\dot{b}_t$ and $A_t = \ddot{b}_t^{-1}$.

Using the definitions in (10.45), we have

$$A_t = 2\exp(\tilde{\theta}_t)\,/\,z_t^2, \qquad x_t = \tilde{\theta}_t + 1 - \exp(\tilde{\theta}_t)\,/\,z_t^2,$$

where we note that A_t is always positive as required. The method of computing the mode based on the two derivatives proceeds as before.

Next we replace $p(\varepsilon_t) = \mathrm{N}(0,1)$ in (9.26) by the Student's t-density (9.23) with $\sigma_\varepsilon^2 = 1$, that is

$$\log p(\varepsilon_t) = \text{constant} - \frac{\nu+1}{2}\log\big(1 + \frac{\varepsilon_t^2}{\nu-2}\big),$$

where $\nu > 2$ is the number of degrees of freedom Hence we obtain the SV model with t-disturbances as discussed in Subsection 9.5.3. We can represent the resulting density for y_t by

$$\log p(y_t|\theta_t) = \text{constant} - \frac{1}{2}\big[\theta_t + (\nu+1)\,\log q_t\big], \qquad q_t = 1 + \exp(-\theta_t)\frac{z_t^2}{\nu-2},$$

for $t = 1, \ldots, n$. For estimating the mode of θ_t we require

$$\dot{p}_t = -\frac{1}{2}\big[1 - (\nu+1)(q_t^{-1} - 1)\big], \qquad \ddot{p}_t = \frac{1}{2}(\nu+1)(q_t^{-1} - 1)q_t^{-1}.$$

Using the definitions in (10.45), we have

$$A_t = 2(\nu+1)^{-1}(\tilde{q}_t - 1)^{-1}\tilde{q}_t^2, \qquad x_t = \tilde{\theta}_t + \tilde{q}_t^2 - \frac{1}{2}A_t,$$

where $\tilde{q}_t$ is q_t evaluated at $\tilde{\theta}_t$. Since $q_t > 1$, all A_t's are positive with probability one and we can proceed with the estimation of the mode.

Other variations of the stochastic volatility model as discussed in Section 9.5 can be considered as well. Estimation of the mode can be carried out similarly for more advanced specifications of the stochastic volatility model. For example, the details of mode estimation for the SV model with leverage effects, as discussed in Subsection 9.5.5, are given by Jungbacker and Koopman (2007).

10.7 Further advances in mode estimation

The basic ideas of mode estimation, together with illustrations, for state space models with a nonlinear non-Gaussian observation equation that depends on a linear Gaussian signal, are presented in Section 10.6. Although many nonlinear non-Gaussian models in practice belong to this class of models, we need to discuss more general cases as well for completeness. We therefore start to derive the general linearisation method based on the state vector for estimating the mode of the state vector. This method is illustrated by a set of examples within the general class of nonlinear non-Gaussian state space models. We close this section by deriving optimality properties of the mode for our class of models.

10.7.1 Linearisation based on the state vector

In Section 10.6 we have considered models where the observation density, conditional on a linear Gaussian signal, is non-Gaussian or nonlinear. The presented method of mode estimation is preferred when it is sufficient to condition on the signal for capturing all nonlinear non-Gaussian features of the model. These developments are also applicable when we need to condition on the state vector rather than the signal. The second-order expansion arguments are still applicable and we can pursue this approach as in the previous section. However we find the linearisation argument for obtaining the mode more insightful, especially when models become more intricate. We therefore present the general linearisation method next.

10.7.2 Linearisation for linear state equations

Consider the non-Gaussian state space model (10.34) but with a linear state equation, that is

$$y_t \sim p(y_t|\alpha_t), \qquad \alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \qquad t = 1, \ldots, n,$$

where the relation between y_t and α_t in density $p(y_t|\alpha_t)$ can be nonlinear and where $p(\alpha_1)$ and $p(\eta_t)$ can be non-Gaussian. We introduce two series of variables $\bar{y}_t$ and $\bar{\alpha}_t$ for $t = 1, \ldots, n$. Let $g(\bar{\alpha}|\bar{y})$ and $g(\bar{\alpha}, \bar{y})$ be the conditional and joint densities generated by the linear Gaussian model (3.1) where observation y_t is replaced by $\bar{y}_t$ and state α_t is replaced by $\bar{\alpha}_t$. Define $\bar{y} = (\bar{y}_1', \ldots, \bar{y}_n')'$ and $\bar{\alpha} = (\bar{\alpha}_1', \ldots, \bar{\alpha}_n')'$. We use notation $\bar{y}_t$ and $\bar{\alpha}_t$ as these quantities are not necessarily equivalent to y_t and α_t, respectively, in our treatment below. All variables x related to (3.1) will be indicated by $\bar{x}$ for $x = Z_t, H_t, T_t, R_t, Q_t, \varepsilon_t, \eta_t, a_1, P_1$ in the same way as for y_t and α_t. Let $p(\alpha|Y_n)$ and $p(\alpha, Y_n)$ be the corresponding densities generated by the general model (10.34) for the observation vector Y_n.

Taking the Gaussian model first, the mode $\hat{\bar{\alpha}}$ is the solution of the vector equation $\partial \log g(\bar{\alpha}|\bar{y})/\partial\bar{\alpha} = 0$. Now $\log g(\bar{\alpha}|\bar{y}) = \log g(\bar{\alpha}, \bar{y}) - \log g(\bar{y})$. Thus the mode is also the solution of the vector equation $\partial \log g(\bar{\alpha}, \bar{y})/\partial\bar{\alpha} = 0$. This version of the equation is easier to manage since $g(\bar{\alpha}, \bar{y})$ has a simple form whereas $g(\bar{\alpha}|\bar{y})$ does not. Since R_t is the linear Gaussian model (3.1) consists of columns of I_m, $\bar{\eta}_t = \bar{R}_t'(\bar{\alpha}_{t+1} - \bar{T}_t\bar{\alpha}_t)$. Assuming that $g(\bar{\alpha}_1) = \mathrm{N}(\bar{a}_1, \bar{P}_1)$, we therefore have

$$\begin{aligned}\log g(\bar{\alpha}, \bar{y}) = \text{constant} &- \frac{1}{2}(\bar{\alpha}_1 - \bar{a}_1)'\bar{P}_1^{-1}(\bar{\alpha}_1 - \bar{a}_1)\\ &- \frac{1}{2}\sum_{t=1}^{n}\left(\bar{\alpha}_{t+1} - \bar{T}_t\bar{\alpha}_t\right)'\bar{R}_t\bar{Q}_t^{-1}\bar{R}_t'(\bar{\alpha}_{t+1} - \bar{T}_t\bar{\alpha}_t)\\ &- \frac{1}{2}\sum_{t=1}^{n}\left(\bar{y}_t - \bar{Z}_t\bar{\alpha}_t\right)'\bar{H}_t^{-1}(\bar{y}_t - \bar{Z}_t\bar{\alpha}_t). \qquad (10.51)\end{aligned}$$

Differentiating with respect to $\bar{\alpha}_t$ and equating to zero gives the equations

$$\begin{aligned}(d_t - 1)\bar{P}_1^{-1}(\bar{\alpha}_1 - \bar{a}_1) - d_t\bar{R}_{t-1}\bar{Q}_{t-1}^{-1}\bar{R}_{t-1}'(\bar{\alpha}_t - \bar{T}_{t-1}\bar{\alpha}_{t-1})&\\ + \bar{T}_t'\bar{R}_t\bar{Q}_t^{-1}\bar{R}_t'(\bar{\alpha}_{t+1} - \bar{T}_t\bar{\alpha}_t) + \bar{Z}_t'\bar{H}_t^{-1}(\bar{y}_t - \bar{Z}_t\bar{\alpha}_t) = 0,& \qquad (10.52)\end{aligned}$$

for $t = 1, \ldots, n$, where $d_1 = 0$ and $d_t = 1$ for $t = 2, \ldots, n$, together with the equation

$$\bar{R}_n\bar{Q}_n\bar{R}_n'(\bar{\alpha}_{n+1} - \bar{T}_n\bar{\alpha}_n) = 0.$$

The solution to these equations is the conditional mode $\hat{\bar{\alpha}}$. Since $g(\bar{\alpha}|\bar{y})$ is Gaussian the mode is equal to the mean so $\hat{\bar{\alpha}}$ can be routinely calculated by the Kalman filter and smoother. We conclude that linear equations of the form

(10.52) can be solved by the Kalman filter and smoother which is efficient computationally.

Assuming that the nonlinear non-Gaussian state space model (10.34) is sufficiently well behaved, the mode $\hat{\alpha}$ of $p(\alpha|Y_n)$ is the solution of the vector equation

$$\frac{\partial \log p(\alpha|Y_n)}{\partial \alpha} = 0$$

and hence, as in the Gaussian case, of the equation

$$\frac{\partial \log p(\alpha, Y_n)}{\partial \alpha} = 0,$$

where

$$\log p(\alpha, Y_n) = \text{constant} + \log p(\alpha_1) + \sum_{t=1}^{n} \left[\log p(\eta_t) + \log p(y_t|\theta_t)\right], \qquad (10.53)$$

with $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$. The mode $\hat{\alpha}$ is the solution of the vector equations

$$\begin{aligned}\frac{\partial \log p(\alpha, Y_n)}{\partial \alpha_t} &= (1 - d_t)\frac{\partial \log p(\alpha_1)}{\partial \alpha_1} + d_t R_{t-1}\frac{\partial \log p(\eta_{t-1})}{\partial \eta_{t-1}} \\ &\quad - T_t' R_t \frac{\partial \log p(\eta_t)}{\partial \eta_t} + \frac{\partial \log p(y_t|\alpha_t)}{\partial \alpha_t} = 0, \qquad (10.54)\end{aligned}$$

for $t = 1, \dots, n$, where, as before, $d_1 = 0$ and $d_t = 1$ for $t = 2, \dots, n$, together with the equation

$$R_n \frac{\partial \log p(\eta_n)}{\partial \eta_n} = 0.$$

We solve these equations by iteration, where at each step we linearise, put the result in the form (10.52) and solve by the Kalman filter and smoother. The final linearised model in the iteration is then the linear Gaussian model with the same conditional mode of $\bar{\alpha}$ given $\bar{y}$ as the non-Gaussian model with the conditional mode α given Y_n.

We first consider the linearisation of the state component in equations (10.54) for a case where the state disturbances η_t are non-Gaussian. Suppose that $\tilde{\eta} = [\tilde{\eta}_1', \dots, \tilde{\eta}_n']'$ is a trial value of $\eta = (\eta_1', \dots, \eta_n')'$ where $\tilde{\eta}_t = R_t'\left(\tilde{\alpha}_{t+1} - T_t\tilde{\alpha}_t\right)$. We shall confine ourselves to the situation where the elements η_{it} of η_t are mutually independent, in other words Q_t is diagonal. Then the state contribution to the conditional mode equations (10.54) is

$$d_t \sum_{i=1}^{r} R_{i,t-1} \frac{\partial \log p(\eta_{i,t-1})}{\partial \eta_{i,t-1}} - T_t' \sum_{i=1}^{r} R_{it} \frac{\partial \log p(\eta_{it})}{\partial \eta_{it}},$$

where we denote the ith column of R_t by R_{it} for $t = 1, \dots, n$ and $i = 1, \dots, r$. Define

$$q_{it} = \left. \frac{\partial \log p(\eta_{it})}{\partial \eta_{it}} \right|_{\eta_t = \tilde{\eta}_t}.$$

The linearised form at $\eta = \tilde{\eta}$ is then given by

$$d_t \sum_{i=1}^{r} R_{i,t-1} q_{i,t-1} \; - \; T_t' \sum_{i=1}^{r} R_{it} q_{it},$$

which has the same form as the state contribution of (10.52) when we set

$$\bar{Q}_t^{-1} = \mathrm{diag}(q_{1t}, \dots, q_{rt}),$$

since $\eta_t = R_t'(\alpha_{t+1} - T_t \alpha_t)$ for $t = 1, \dots, n$. All other state variables $\bar{x}$ are set equal to x for $x = Z_t, H_t, T_t, R_t, \dots$. In the iterative estimation of $\hat{\alpha}$ the Kalman filter and smoother can be used to update the trial value $\tilde{\alpha}$ and, consequently, $\tilde{\eta}$.

10.7.3 Linearisation for nonlinear models

Consider the non-Gaussian state space model (10.34) where the density of y_t is subject to a nonlinear signal, $p(y_t|\theta_t) = p(y_t|\alpha_t)$, where

$$\theta_t = Z_t(\alpha_t), \qquad t = 1, \dots, n. \tag{10.55}$$

Our objective is to find an approximating linear Gaussian model with the same conditional mode of $\bar{\alpha}$ given $\bar{y}$ as the nonlinear model. We do this by a technique which is slightly different, though simpler, than that used for the non-Gaussian models. The basic idea is to linearise the observation and state equations (10.34) and (10.55) directly, which immediately delivers an approximating linear Gaussian model. We then iterate to ensure that this approximating model has the same conditional mode as the original nonlinear model.

Taking first the nonlinear signal (10.55), let $\tilde{\alpha}_t$ be a trial value of α_t. Expanding about $\tilde{\alpha}_t$ gives approximately

$$Z_t(\alpha_t) = Z_t(\tilde{\alpha}_t) + \dot{Z}_t(\tilde{\alpha}_t)(\alpha_t - \tilde{\alpha}_t),$$

where $\dot{Z}(\alpha_t) = \partial Z_t(\alpha_t)/\partial \alpha_t'$. From an approximation to (10.55) we obtain

$$y_t = \bar{d}_t + \dot{Z}_t(\tilde{\alpha}_t)\alpha_t + \varepsilon_t, \qquad \bar{d}_t = Z_t(\tilde{\alpha}_t) - \dot{Z}_t(\tilde{\alpha}_t)\tilde{\alpha}_t, \tag{10.56}$$

which is the linear observation equation in the mean adjusted form; see Subsection 4.3.3. Similarly, if we expand the state updating function in (10.34) about $\tilde{\alpha}_t$ we obtain approximately

$$T_t(\alpha_t) = T_t(\tilde{\alpha}_t) + \dot{T}_t(\tilde{\alpha}_t)(\alpha_t - \tilde{\alpha}_t),$$

where $\dot{T}(\alpha_t) = \partial T_t(\alpha_t)/\partial \alpha_t'$. Thus we obtain the linearised relation

$$\alpha_{t+1} = \bar{c}_t + \dot{T}_t(\tilde{\alpha}_t)\alpha_t + R_t(\tilde{\alpha}_t)\eta_t, \qquad \bar{c}_t = T_t(\tilde{\alpha}_t) - \dot{T}_t(\tilde{\alpha}_t)\tilde{\alpha}_t. \tag{10.57}$$

We approximate this nonlinear model at its conditional mode by the linear Gaussian model with mean adjustments as discussed in Subsection 4.3.3 with the modified form

$$\begin{aligned} \bar{y}_t &= \bar{d}_t + \bar{Z}_t\alpha_t + \bar{\varepsilon}_t, & \bar{\varepsilon}_t &\sim \mathrm{N}(0, \bar{H}_t), \\ \alpha_{t+1} &= \bar{c}_t + \bar{T}_t\alpha_t + \bar{R}_t\eta_t, & \bar{\eta}_t &\sim \mathrm{N}(0, \bar{Q}_t), \end{aligned} \tag{10.58}$$

where

$$\bar{Z}_t = \dot{Z}_t(\tilde{\alpha}_t), \quad \bar{H}_t = H_t(\tilde{\alpha}_t), \quad \bar{T}_t = \dot{T}_t(\tilde{\alpha}_t), \quad \bar{R}_t = R_t(\tilde{\alpha}_t), \quad \bar{Q}_t = Q_t(\tilde{\alpha}_t),$$

for $t = 1, \ldots, n$. The Kalman filter of the form (4.25) in Subsection 4.3.3 can be applied to model (10.58). We use the output of the Kalman filter (4.26) to define a new $\tilde{\alpha}_t$ which gives a new approximating model (10.58), and we continue to iterate as in Subsection 10.6.3 until convergence is achieved. Denote the resulting value of α by $\hat{\alpha}$.

10.7.4 Linearisation for multiplicative models

Shephard (1994b) considered the multiplicative trend and seasonal model with additive Gaussian observation noise. We consider here a simple version of this model with the trend modelled as a local level and the seasonal given by a single trigonometric term such as given in (3.6) with $s = 3$. We have

$$y_t = \mu_t\gamma_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big),$$

with

$$\alpha_{t+1} = \begin{pmatrix} \mu_{t+1} \\ \gamma_{t+1} \\ \gamma_{t+1}^* \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\lambda & \sin\lambda \\ 0 & -\sin\lambda & \cos\lambda \end{bmatrix} \alpha_t + \begin{pmatrix} \eta_t \\ \omega_t \\ \omega_t^* \end{pmatrix},$$

and $\lambda = 2\pi/3$. It follows that $Z_t(\alpha_t) = \mu_t\gamma_t$ and $\dot{Z}_t(\alpha_t) = (\gamma_t, \mu_t, 0)$ which lead us to the approximating model

$$\tilde{y}_t = (\tilde{\gamma}_t, \tilde{\mu}_t, 0)\alpha_t + \varepsilon_t,$$

where $\tilde{y}_t = y_t + \tilde{\mu}_t\tilde{\gamma}_t$.

Another example of a multiplicative model that we consider is

$$y_t = \mu_t\varepsilon_t, \qquad \mu_{t+1} = \mu_t\xi_t,$$

where ε_t and ξ_t are mutually and serially uncorrelated Gaussian disturbance terms. For the general model (9.36) and (9.37) we have $\alpha_t = (\mu_t, \varepsilon_t, \xi_t)'$, $\eta_t =$

$(\varepsilon_{t+1}, \xi_{t+1})'$ $Z_t(\alpha_t) = \mu_t \varepsilon_t$, $H_t = 0$, $T_t(\alpha_t) = (\mu_t \xi_t, 0, 0)'$, $R_t = [0, I_2]'$ and Q_t is a 2×2 diagonal matrix. It follows that

$$\dot{Z}_t(\alpha_t) = (\varepsilon_t, \mu_t, 0), \qquad \dot{T}_t(\alpha_t) = \begin{bmatrix} \xi_t & 0 & \mu_t \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The approximating model (10.56) and (10.57) reduces to

$$\tilde{y}_t = \tilde{\varepsilon}_t \mu_t + \tilde{\mu}_t \varepsilon_t, \qquad \mu_{t+1} = -\tilde{\mu}_t \tilde{\xi}_t + \tilde{\xi}_t \mu_t + \tilde{\mu}_t \xi_t,$$

with $\tilde{y}_t = y_t + \tilde{\mu}_t \tilde{\varepsilon}_t$. Thus the Kalman filter and smoother can be applied to an approximating time-varying local level model with state vector $\alpha_t = \mu_t$.

10.7.5 An optimal property for the mode

We will emphasise in the next chapter the use of the mode $\hat{\alpha}$ of $p(\alpha|Y_n)$ to obtain a linear approximating model which we use for simulation. If, however, the sole object of the investigation was to estimate α, then $\hat{\alpha}$ could be used for the purpose without recurse to simulation; indeed, this was the estimator used by Durbin and Koopman (1992) and an approximation to it was used by Fahrmeir (1992).

The property that the conditional mode is the most probable value of the state vector given the observations can be regarded as an optimality property; we now consider a further optimality property possessed by the conditional mode. To find it we examine the analogous situation in maximum likelihood estimation. The maximum likelihood estimate of a parameter ψ is the most probable value of it given the observations and is well known to be asymptotically efficient. To develop a finite-sample property analogous to asymptotic efficiency, Godambe (1960) and Durbin (1960) introduced the idea of unbiased estimating equations and Godambe showed that the maximum likelihood estimate of scalar ψ is the solution to an unbiased estimating equation which has a minimum variance property. This can be regarded as a finite-sample analogue of asymptotic efficiency. The extension to multidimensional ψ was indicated by Durbin (1960). Since that time there have been extensive developments of this basic idea, as can be seen from the collection of papers edited by Basawa, Godambe and Taylor (1997). Following Durbin (1997), we now develop a minimum-variance unbiased estimating equation property for the conditional mode estimate $\hat{\alpha}$ of the random vector α.

If α^* is the unique solution for α of the $mn \times 1$ vector equation $H(\alpha, Y_n) = 0$ and if $\mathrm{E}[H(\alpha, Y_n)] = 0$, where expectation is taken with respect to the joint density $p(\alpha, Y_n)$, we say that $H(\alpha, Y_n) = 0$ is an *unbiased estimating equation*. It is obvious that the equation can be multiplied through by an arbitrary nonsingular matrix and still give the same solution α^*. We therefore standardise $H(\alpha, Y_n)$ in the way that is usual in estimating equation theory and multiply it by $[\mathrm{E}\{\dot{H}(\alpha, Y_n)\}]^{-1}$, where $\dot{H}(\alpha, Y_n) = \partial H(\alpha, Y_n)/\partial\alpha'$, and then seek a minimum variance property for the resulting function $h(\alpha, Y_n) = [\mathrm{E}\{\dot{H}(\alpha, Y_n)\}]^{-1}H(\alpha, Y_n)$.

Let

$$\mathrm{Var}[h(\alpha, Y_n)] = \mathrm{E}[h\,(\alpha, Y_n)\,h\,(\alpha, Y_n)'],$$
$$\mathcal{J} = \mathrm{E}\left[\frac{\partial \log p(\alpha, Y_n)}{\partial\alpha}\frac{\partial \log p(\alpha, Y_n)}{\partial\alpha'}\right].$$

Under mild conditions that are likely to be satisfied in many practical cases, Durbin (1997) showed that $\mathrm{Var}\,[h\,(\alpha, Y_n)] - \mathcal{J}^{-1}$ is non-negative definite. If this is a zero matrix we say that the corresponding equation $H(\alpha, Y_n) = 0$ is an *optimal estimating equation.* Now take $H(\alpha, Y_n) = \partial \log p(\alpha, Y_n)/\partial\alpha$. Then $\mathrm{E}[\dot{H}(\alpha, Y_n)] = -\mathcal{J}$, so $h(\alpha, Y_n) = -\mathcal{J}^{-1}\partial \log p(\alpha, Y_n)/\partial\alpha$. Thus $\mathrm{Var}[h(\alpha, Y_n)] = \mathcal{J}^{-1}$ and consequently the equation $\partial \log p(\alpha, Y_n)/\partial\alpha = 0$ is optimal. Since $\hat{\alpha}$ is the solution of this, it is the solution of an optimal estimating equation. In this sense the conditional mode has an optimality property analogous to that of maximum likelihood estimates of fixed parameters in finite samples.

We have assumed above that there is a single mode and the question arises whether multimodality will create complications. If multimodality is suspected it can be investigated by using different starting points and checking whether iterations from them converge to the same mode. In none of the cases we have examined has multimodality of $p(\alpha|Y_n)$ caused any difficulties. For this reason we regard it as unlikely that it will give rise to problems in routine time series analysis. If, however, multimodality were to occur in a particular case, we would suggest fitting a linear Gaussian model to the data at the outset and using this to define the first importance density $g_1(\eta|Y_n)$ and conditional joint density $g_1(\eta, Y_n)$. Simulation is employed to obtain a first estimate $\tilde{\eta}^{(1)}$ of $\mathrm{E}(\eta|Y_n)$ and from this a first estimate $\tilde{\theta}_t^{(1)}$ of θ_t is calculated for $t = 1, \ldots, n$. Now linearise the true densities at $\tilde{\eta}^{(1)}$ or $\tilde{\theta}_t^{(1)}$ to obtain a new approximating linear Gaussian model which defines a new $g(\eta|Y_n)$, $g_2(\eta|Y_n)$, and a new $g(\eta, Y_n)$, $g_2(\eta, Y_n)$. Simulation using these gives a new estimate $\tilde{\eta}^{(2)}$ of $\mathrm{E}(\eta|Y_n)$. This iterative process is continued until adequate convergence is achieved. We emphasise, however, that it is not necessary for the final value of α at which the model is linearised to be a precisely accurate estimate of either the mode or the mean of $p(\alpha|Y_n)$. The only way that the choice of the value of α used as the basis for the simulation affects the final estimate $\hat{x}$ is in the variances due to simulation as we shall show later. Where necessary, the simulation sample size can be increased to reduce these error variances to any required extent. It will be noted that we are basing these iterations on the mean, not the mode. Since the mean, when it exists, is unique, no question of 'multimeanality' can arise.

10.8 Treatments for heavy-tailed distributions

In this section we consider both approximate and exact treatments of linear state space models with disturbances from heavy-tailed distributions. First we apply linearisation techniques based on the first derivative only for different model specifications. This shows that the general methodology can lead to practical methods for treating outliers and breaks in time series. The methods lead to mode estimates. We further discuss simulation treatments for heavy-tailed models which are relatively simple and lead to exact estimation methods subject to simulation error.

10.8.1 Mode estimation for models with heavy-tailed densities

In a signal plus noise model, $y_t = \theta_t + \varepsilon_t$ where the signal θ_t is linear Gaussian and the noise distribution for ε_t has a heavy-tailed density, we can estimate the mode using the first two derivatives or using the first derivative only. For example, consider the logdensity of the Student's t-distribution as specified in (9.23). To estimate the mode of θ_t based on (10.44) we require expressions for A_t, the tth element of diagonal matrix A, and x_t, the tth element of vector x, where A and x are given by (10.45). The variables A_t and x_t rely on $\dot{p}(\varepsilon_t) = \dot{p}(y_t|\theta_t)$ and $\ddot{p}(\varepsilon_t) = \ddot{p}(y_t|\theta_t)$ which are given by

$$\dot{p}(\varepsilon_t) = (\nu+1)s_t^{-1}\tilde{\varepsilon}_t, \qquad \ddot{p}(\varepsilon_t) = (\nu+1)s_t^{-1}\left[2s_t^{-1}\tilde{\varepsilon}_t - 1\right],$$

where $s_t = (\nu-2)\sigma_\varepsilon^2 + \tilde{\varepsilon}_t^2$. Since we cannot rule out a positive value for $\ddot{p}(\varepsilon_t)$, the t-density $p(\varepsilon_t)$ is not logconcave in θ_t and the variance A_t can become negative. However, the method of mode estimation is still applicable in this case; see Jungbacker and Koopman (2007). When we prefer to estimate the mode using the first derivative only as in (10.49), we can adopt the linear Gaussian signal plus noise model with its observation variance given by

$$A_t = (\nu+1)^{-1}s_t,$$

for $t = 1, \ldots, n$. We proceed by applying the Kalman filter and smoother to obtain a new smooth estimate of θ_t.

Similar computations can be adopted for other heavy-tailed densities. In the case of the mixture of normals model with density (9.24) we obtain,

$$\log p(\varepsilon_t) = \log\left\{\lambda^*\, e_t(\sigma_\varepsilon^2) \;+\; [1-\lambda^*]\, e_t(\sigma_\varepsilon^2\chi)\right\},$$

where $e_t(z) = \exp(-\frac{1}{2}\varepsilon_t^2/z)\,/\,\sqrt{2\pi z}$, with $1-\lambda^*$ as the proportion of outliers and χ as the multiplier of the variance for the outliers. For the computation of the mode using the first derivative only, we require

$$A_t = p(\tilde{\varepsilon}_t)\left\{\lambda^*\,\sigma_\varepsilon^{-2}\,\tilde{e}_t(\sigma_\varepsilon^2) \;+\; [1-\lambda^*]\,(\sigma_\varepsilon^2\chi)^{-1}\,\tilde{e}_t(\sigma_\varepsilon^2\chi)\right\}^{-1}$$

where $\tilde{\varepsilon}_t$ is a particular value of ε_t and $\tilde{e}_t(z)$ is $e_t(z)$ evaluated at $\varepsilon_t = \tilde{\varepsilon}_t$ for any $z > 0$.

Finally, in the case of the general error density (9.25) we obtain,

$$\log p(\varepsilon_t) = \text{constant} - c(\ell) \left| \frac{\varepsilon_t}{\sigma_\varepsilon} \right|^{\ell},$$

for coefficient $1 < \ell < 2$ and with $c(\ell)$ as a known function of ℓ. For given values $\varepsilon_t = \tilde{\varepsilon}_t$, we can compute the model using the first derivative only by computing

$$A_t = \text{sign}(\tilde{\varepsilon}_t) \frac{\tilde{\varepsilon}_t \sigma_\varepsilon}{c(\ell)\,\ell} \left| \frac{\tilde{\varepsilon}_t}{\sigma_\varepsilon} \right|^{1-\ell}.$$

All A_t's are positive with probability one and we can proceed with the estimation of the mode as before.

10.8.2 Mode estimation for state errors with t-distribution

As an illustration of mode estimation for linear models with a t-distribution, we consider the local level model (2.3) but with a t-distribution for the state error term η_t. We obtain

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t, \qquad && \varepsilon_t \sim \text{N}\big(0, \sigma_\varepsilon^2\big), \\ \alpha_{t+1} &= \alpha_t + \eta_t, \qquad && \eta_t \sim t_\nu, \end{aligned}$$

and we assume that $\alpha_1 \sim \text{N}(0, \kappa)$ with $\kappa \to \infty$. By adopting the same arguments for linearisation using one derivative only, we obtain

$$A_t^* = (\nu + 1)^{-1} s_t^*,$$

where $s_t^* = (\nu - 2)\sigma_\eta^2 + \tilde{\eta}_t^2$, for $t = 1, \ldots, n$. Starting with initial values for $\tilde{\eta}_t$, we compute A_t^* and apply the Kalman filter and disturbance smoother to the approximating Gaussian local level model with

$$y_t = \alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = \alpha_t + \eta_t, \qquad \eta_t \sim \text{N}(0, A_t^*).$$

New values for the smoothed estimates $\tilde{\eta}_t$ are used to compute new values for A_t^* until convergence to $\hat{\eta}_t$. When we assume that both disturbances ε_t and η_t in the local level model (2.3) are generated by t-distributions, we can obtain the mode by computing both A_t and A_t^* and adopt them as variances for the two corresponding disturbances in the linear Gaussian local level model.

10.8.3 A simulation treatment for t-distribution model

In some cases it is possible to construct simulations by using antithetic variables without importance sampling. For example, it is well-known that if a random variable u_t has the standard t-distribution with ν degrees of freedom then u_t has the representation

$$u_t = \frac{\nu^{1/2}\varepsilon_t^*}{c_t^{1/2}}, \qquad \varepsilon_t^* \sim \mathrm{N}(0,1), \qquad c_t \sim \chi^2(\nu), \qquad \nu > 2, \tag{10.59}$$

where ε_t^* and c_t are independent. In the case where ν is not an integer we take $\frac{1}{2}c_t$ as a gamma variable with parameter $\frac{1}{2}\nu$. It follows that if we consider the case where ε_t is univariate and we take ε_t in model (9.4) to have logdensity (9.23) then ε_t has the representation

$$\varepsilon_t = \frac{(\nu-2)^{1/2}\sigma_\varepsilon\varepsilon_t^*}{c_t^{1/2}}, \tag{10.60}$$

where ε_t^* and c_t are as in (10.59). Now take $\varepsilon_1^*, \ldots, \varepsilon_n^*$ and $c_1, \ldots, c_n$ to be mutually independent. Then conditional on $c_1, \ldots, c_n$ fixed, model (9.4) and (9.2), with $\eta_t \sim \mathrm{N}(0, Q_t)$, is a linear Gaussian model with $H_t = \mathrm{Var}(\varepsilon_t) = (\nu-2)\sigma_\varepsilon^2 c_t^{-1}$. Put $c = (c_1, \ldots, c_n)'$. We now show how to estimate the conditional means of functions of the state using simulation samples from the distribution of c.

Suppose first that α_t is generated by the linear Gaussian model $\alpha_{t+1} = T_t\alpha_t + R_t\eta_t$, $\eta_t \sim \mathrm{N}(0, Q_t)$, and that as in (11.12) we wish to estimate

$$\begin{aligned}
\bar{x} &= \mathrm{E}[x^*(\eta)|y] \\
&= \int x^*(\eta)p(c,\eta|y)\,dc\,d\eta \\
&= \int x^*(\eta)p(\eta|c,y)p(c|y)\,dc\,d\eta \\
&= \int x^*(\eta)p(\eta|c,y)p(c,y)p(y)^{-1}dc\,d\eta \\
&= p(y)^{-1}\int x^*(\eta)p(\eta|c,y)p(y|c)p(c)\,dc\,d\eta.
\end{aligned} \tag{10.61}$$

For given c, the model is linear and Gaussian. Let

$$\bar{x}(c) = \int x^*(\eta)p(\eta|c,y)\,d\eta.$$

For many cases of interest, $\bar{x}(c)$ is easily calculated by the Kalman filter and smoother, as in Chapters 4 and 5; to begin with, let us restrict attention to these cases. We have

$$p(y) = \int p(y,c)\,dc = \int p(y|c)p(c)\,dc,$$

where $p(y|c)$ is the likelihood given c which is easily calculated by the Kalman filter as in 7.2. Denote expectation with respect to density $p(c)$ by E_c Then from (10.61),

$$\bar{x} = \frac{\mathrm{E}_c[\bar{x}(c)p(y|c)]}{\mathrm{E}_c[p(y|c)]}. \tag{10.62}$$

We estimate this by simulation. Independent simulation samples $c^{(1)}, c^{(2)}, \ldots$ of c are easily obtained since c is a vector of independent χ^2_ν variables. We suggest that antithetic values of χ^2_ν are employed for each element of c, either in balanced pairs or balanced sets of four as described in Subsection 11.4.3. Suppose that values $c^{(1)}, \ldots, c^{(N)}$ have been selected. Then estimate $\bar{x}$ by

$$\hat{x} = \frac{\sum_{i=1}^N \bar{x}\big(c^{(i)}\big)p\big(y|c^{(i)}\big)}{\sum_{i=1}^N p\big(y|c^{(i)}\big)}. \tag{10.63}$$

When $\bar{x}(c)$ cannot be computed by the Kalman filter and smoother we first draw a value of $c^{(i)}$ as above and then for the associated linear Gaussian model that is obtained when this value $c^{(i)}$ is fixed we draw a simulated value $\eta^{(i)}$ of η using the simulated smoother of Section 4.9, employing antithetics independently for both $c^{(i)}$ and $\eta^{(i)}$. The value $x^*(\eta^{(i)})$ is then calculated for each $\eta^{(i)}$. If there are N pairs of values $c^{(i)}, \eta^{(i)}$ we estimate $\bar{x}$ by

$$\hat{x}^* = \frac{\sum_{i=1}^N x^*\big(\eta^{(i)}\big)p\big(y|c^{(i)}\big)}{\sum_{i=1}^N p\big(y|c^{(i)}\big)}. \tag{10.64}$$

Since we now have sampling variation arising from the drawing of values of η as well as from drawing values of c, the variance of $\hat{x}^*$ will be larger than of $\hat{x}$ for a given value of N. We present formulae (10.63) and (10.64) at this point for expository convenience in advance of the general treatment of analogous formulae in Section 11.5.

Now consider the case where the error term ε_t in the observation equation is $\mathrm{N}(0, \sigma^2_\varepsilon)$ and where the elements of the error vector η_t in the state equation are independently distributed as Student's t. For simplicity assume that the number of degrees of freedom in these t-distributions are all equal to ν, although there is no difficulty in extending the treatment to the case where some of the degrees of freedom differ or where some elements are normally distributed. Analogously to (10.60) we have the representation

$$\eta_{it} = \frac{(\nu-2)^{1/2}\sigma_{\eta i}\eta^*_{it}}{c_{it}^{1/2}}, \qquad \eta^*_{it} \sim \mathrm{N}(0,1), \qquad c_{it} \sim \chi^2_\nu, \qquad \nu > 2, \tag{10.65}$$

for $i = 1, \ldots, r$ and $t = 1, \ldots, n$, where $\sigma^2_{\eta i} = \mathrm{Var}(\eta_{it})$. Conditional on $c_{11}, \ldots, c_{rn}$ held fixed, the model is linear and Gaussian with $H_t = \sigma^2_\varepsilon$ and

$\eta_t \sim \mathrm{N}(0, Q_t)$ where $Q_t = \mathrm{diag}[(\nu - 2)\sigma^2_{\eta 1} c^{-1}_{1t}, \ldots, (\nu - 2)\sigma^2_{\eta r} c^{-1}_{rt}]$. Formulae (10.63) and (10.64) remain valid except that $c^{(i)}$ is now a vector with r elements. The extension to the case where both ε_t and elements of η_t have t-distributions is straightforward.

The idea of using representation (10.59) for dealing with disturbances with t-distributions in the local level model by means of simulation was proposed by Shephard (1994b) in the context of MCMC simulation.

10.8.4 A simulation treatment for mixture of normals model

An alternative to the t-distribution for representing error distributions with heavy tails is to employ the Gaussian mixture density (9.24), which for univariate ε_t we write in the form

$$p(\varepsilon_t) = \lambda^* \mathrm{N}\big(0, \sigma^2_\varepsilon\big) + (1 - \lambda^*)\mathrm{N}\big(0, \chi\sigma^2_\varepsilon\big), \qquad 0 < \lambda^* < 1. \tag{10.66}$$

It is obvious that values of ε_t with this density can be realised by means of a two-stage process in which we first select the value of a binomial variable b_t such that $\Pr(b_t = 1) = \lambda^*$ and $\Pr(b_t = 0) = 1 - \lambda^*$, and then take $\varepsilon_t \sim \mathrm{N}(0, \sigma^2_\varepsilon)$ if $b_t = 1$ and $\varepsilon_t \sim \mathrm{N}(0, \chi\sigma^2_\varepsilon)$ if $b_t = 0$. Assume that the state vector α_t is generated by the linear Gaussian model $\alpha_{t+1} = T_t\alpha_t + R_t\eta_t$, $\eta_t \sim \mathrm{N}(0, Q_t)$. Putting $b = (b_1, \ldots, b_n)'$, it follows that for b given, the state space model is linear and Gaussian. We can therefore employ the same approach for the mixture distribution that we used for the t-distribution in the previous subsection, giving as in (10.61),

$$\bar{x} = p(y)^{-1} M^{-1} \sum_{j=1}^{M} \int x^*(\eta) p\big(\eta | b_{(j)}, y\big) p\big(y | b_{(j)}\big) p\big(b_{(j)}\big)\, d\eta, \tag{10.67}$$

where $b_{(1)}, \ldots, b_{(M)}$ are the $M = 2^n$ possible values of b. Let

$$\bar{x}(b) = \int x^*(\eta) p(\eta | b, y)\, d\eta,$$

and consider cases where this can be calculated by the Kalman filter and smoother. Denote expectation over the distribution of b by E_b. Then

$$p(y) = M^{-1} \sum_{j=1}^{M} p\big(y | b_{(j)}\big) p\big(b_{(j)}\big) = \mathrm{E}_b[p(y|b)],$$

and analogously to (10.62) we have,

$$\bar{x} = \frac{\mathrm{E}_b[\bar{x}(b) p(y|b)]}{\mathrm{E}_b[p(y|b)]}. \tag{10.68}$$

We estimate this by simulation. A simple way to proceed is to choose a sequence $b^{(1)}, \ldots, b^{(N)}$ of random values of b and then estimate $\bar{x}$ by

$$\hat{x} = \frac{\sum_{i=1}^{N} \bar{x}\big(b^{(i)}\big) p\big(y|b^{(i)}\big)}{\sum_{i=1}^{N} p\big(y|b^{(i)}\big)}. \tag{10.69}$$

Variability in this formula arises only from the random selection of b. To construct antithetic variables for the problem we consider how this variability can be restricted while preserving correct overall probabilities. We suggest the following approach. Consider the situation where the probability $1-\lambda^*$ in (10.66) of taking $\mathrm{N}(0, \chi\sigma_\varepsilon^2)$ is small. Take $1 - \lambda^* = 1/B$ where B is an integer, say $B = 10$ or 20. Divide the simulation sample of values of b into K blocks of B, with $N = KB$. Within each block, and for each $t = 1, \ldots, n$, choose integer j randomly from 1 to B, put the jth value in the block as $b_t = 0$ and the remaining $B - 1$ values in the block as $b_t = 1$. Then take $b^{(i)} = (b_1, \ldots, b_n)'$ with $b_1, \ldots, b_n$ defined in this way for $i = 1, \ldots, N$ and use formula (10.69) to estimate $\bar{x}$. With this procedure we have ensured that for each i, $\Pr(b_t = 1) = \lambda^*$ as desired, with b_s and b_t independent for $s \neq t$, while enforcing balance in the sample by requiring that within each block b_t has exactly $B - 1$ values of 1 and one value of 0. Of course, choosing integers at random from 1 to B is a much simpler way to select a simulation sample than using the simulation smoother.

The restriction of B to integer values is not a serious drawback since the results are insensitive to relatively small variations in the value of λ^*, and in any case the value of λ^* is normally determined on a trial-and-error basis. It should be noted that for purposes of estimating mean square errors due to simulation, the numerator and denominator of (10.69) should be treated as composed of M independent values.

The idea of using the binomial representation of (10.66) in MCMC simulation for the local level model was proposed by Shephard (1994b).

11 Importance sampling for smoothing

11.1 Introduction

In this chapter we develop the methodology of importance sampling based on simulation for the analysis of observations from the non-Gaussian and nonlinear models that we have specified in Chapter 9. Unlike the treatment of linear models in Chapters 2 and 4, we deal with smoothing first and leave filtering for Chapter 12. The main reason is that filtering is dealt with by the method of particle filtering which is based on importance sampling methods such as those discussed in this chapter. We show that importance sampling methods can be adopted for the estimation of functions of the state vector and the estimation of the error variance matrices of the resulting estimates. We shall also develop estimates of conditional densities, distribution functions and quantiles of interest, given the observations. Of key importance is the method of estimating unknown parameters by maximum likelihood. The methods are based on standard ideas in simulation methodology and, in particular, importance sampling. In this chapter we will develop the basic ideas of importance sampling that we employ in our methodology for nonlinear non-Gaussian state space models. Details of applications to particular models will be given in later sections.

Importance sampling was introduced as early as Kahn and Marshall (1953) and Marshall (1956) and was described in books by Hammersley and Handscomb (1964, Section 5.4) and Ripley (1987, Chapter 5). It was first used in econometrics by Kloek and Van Dijk (1978) in their work on computing posterior densities.

The general model of interest in this chapter is given by

$$y_t \sim p(y_t|\alpha_t), \qquad \alpha_{t+1} = T_t(\alpha_t) + R_t\eta_t, \qquad \eta_t \sim p(\eta_t), \tag{11.1}$$

where both $p(y_t|\alpha_t)$ and $p(\eta_t)$, for $t = 1, \dots, n$, can be non-Gaussian densities. We also give attention to the special case of a signal model with a linear Gaussian state equation, that is,

$$y_t \sim p(y_t|\theta_t), \qquad \alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \qquad \eta_t \sim \mathrm{N}(0, Q_t), \tag{11.2}$$

for $t = 1, \dots, n$, where $\theta_t = Z_t\alpha_t$ and R_t is a selection matrix with $R_t'R_t = I_r$ and r is the number of disturbances in η_t; see Table 4.1. Denote the stacked vectors

$(\alpha_1', \ldots, \alpha_{n+1}')'$, $(\theta_1', \ldots, \theta_n')'$ and $(y_1', \ldots, y_n')'$ by α, θ and Y_n, respectively. In order to keep the exposition simple we shall assume in this section and the next section that the initial density $p(\alpha_1)$ is nondegenerate and known. The case where some of the elements of α_1 are diffuse will be considered in Subsection 11.4.4.

We mainly focus on the estimation of the conditional mean

$$\bar{x} = \mathrm{E}[x(\alpha)|Y_n] = \int x(\alpha)p(\alpha|Y_n)\, d\alpha, \tag{11.3}$$

of an arbitrary function $x(\alpha)$ of α given the observation vector Y_n. This formulation includes estimates of quantities of interest such as the mean $\mathrm{E}(\alpha_t|Y_n)$ of the state vector α_t given Y_n and its conditional variance matrix $\mathrm{Var}(\alpha_t|Y_n)$; it also includes estimates of the conditional density and distribution function of $x(\alpha)$ given Y_n when $x(\alpha)$ is scalar. The conditional density $p(\alpha|Y_n)$ depends on an unknown parameter vector ψ, but in order to keep the notation simple we shall not indicate this dependence explicitly in this chapter; the estimation of ψ is considered in Section 11.6.

In theory, we could draw a random sample of values from the distribution with density $p(\alpha|Y_n)$ and estimate $\bar{x}$ by the sample mean of the corresponding values of $x(\alpha)$. In practice, however, since explicit expressions are not available for $p(\alpha|Y_n)$ for the models of Chapter 9, this idea is not feasible. Instead, we seek a density as close to $p(\alpha|Y_n)$ as possible for which random draws are available, and we sample from this, making an appropriate adjustment to the integral in (11.3). This technique is called *importance sampling* and the density is referred to as the *importance density.* The techniques we shall describe will be based on Gaussian importance densities since these are available for the problems we shall consider and they work well in practice. We shall use the generic notation $g(\cdot)$, $g(\cdot,\cdot)$ and $g(\cdot|\cdot)$ for marginal, joint and conditional densities, respectively.

11.2 Basic ideas of importance sampling

Consider model (11.1) and let $g(\alpha|Y_n)$ be an importance density which is chosen to resemble $p(\alpha|Y_n)$ as closely as is reasonably possible while being easy to sample from; we have from (11.3),

$$\bar{x} = \int x(\alpha)\frac{p(\alpha|Y_n)}{g(\alpha|Y_n)}g(\alpha|Y_n)\, d\alpha = \mathrm{E}_g\left[x(\alpha)\frac{p(\alpha|Y_n)}{g(\alpha|Y_n)}\right], \tag{11.4}$$

where E_g denotes expectation with respect to the importance density $g(\alpha|Y_n)$. For the models of Chapter 9, $p(\alpha|Y_n)$ and $g(\alpha|Y_n)$ are complicated algebraically, whereas the corresponding joint densities $p(\alpha, Y_n)$ and $g(\alpha, Y_n)$ are straightforward. We therefore put $p(\alpha|Y_n) = p(\alpha, Y_n)/p(Y_n)$ and $g(\alpha|Y_n) = g(\alpha, Y_n)/g(Y_n)$ in (11.4), giving

$$\bar{x} = \frac{g(Y_n)}{p(Y_n)}\mathrm{E}_g\left[x(\alpha)\frac{p(\alpha, Y_n)}{g(\alpha, Y_n)}\right]. \tag{11.5}$$

Putting $x(\alpha) = 1$ in (11.5) we have

$$1 = \frac{g(Y_n)}{p(Y_n)} \mathrm{E}_g \left[\frac{p(\alpha, Y_n)}{g(\alpha, Y_n)} \right], \tag{11.6}$$

and effectively obtain an expression for the observation density

$$p(Y_n) = g(Y_n)\, \mathrm{E}_g \left[\frac{p(\alpha, Y_n)}{g(\alpha, Y_n)} \right]. \tag{11.7}$$

Taking the ratio of (11.5) and (11.6) gives

$$\bar{x} = \frac{\mathrm{E}_g[x(\alpha) w(\alpha, Y_n)]}{\mathrm{E}_g[w(\alpha, Y_n)]}, \quad \text{where} \quad w(\alpha, Y_n) = \frac{p(\alpha, Y_n)}{g(\alpha, Y_n)}. \tag{11.8}$$

In model (11.1)

$$p(\alpha, Y_n) = p(\alpha_1) \prod_{t=1}^{n} p(\eta_t) p(y_t|\alpha_t), \tag{11.9}$$

where $\eta_t = R_t' \left[\alpha_{t+1} - T_t(\alpha_t) \right]$ for $t = 1, \ldots, n$, since $R_t' R_t = I^r$.

Expression (11.8) provides the basis for importance sampling. We could in principle obtain a Monte Carlo estimate $\hat{x}$ of $\bar{x}$ in the following way. Choose a series of independent draws $\alpha^{(1)}, \ldots, \alpha^{(N)}$ from the distribution with density $g(\alpha|Y_n)$ and take

$$\hat{x} = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i}, \quad \text{where} \quad x_i = x\big(\alpha^{(i)}\big) \quad \text{and} \quad w_i = w\big(\alpha^{(i)}, Y_n\big). \tag{11.10}$$

Since the draws are independent, the law of large numbers applies and under assumptions which are usually satisfied in cases of practical interest, $\hat{x}$ converges to $\bar{x}$ probabilistically as $N \to \infty$.

An important special case is where the observations are non-Gaussian but the state equation is linear Gaussian, that is, where we consider model (11.2). We then have $p(\alpha) = g(\alpha)$ so

$$\frac{p(\alpha, Y_n)}{g(\alpha, Y_n)} = \frac{p(\alpha) p(Y_n|\alpha)}{g(\alpha) g(Y_n|\alpha)} = \frac{p(Y_n|\alpha)}{g(Y_n|\alpha)} = \frac{p(Y_n|\theta)}{g(Y_n|\theta)},$$

where θ is the stacked vector of the signals $\theta_t = Z_t \alpha_t$ for $t = 1, \ldots, n$. Thus (11.8) becomes the simpler formula

$$\bar{x} = \frac{\mathrm{E}_g[x(\alpha) w^*(\theta, Y_n)]}{\mathrm{E}_g[w^*(\theta, Y_n)]}, \quad \text{where} \quad w^*(\theta, Y_n) = \frac{p(Y_n|\theta)}{g(Y_n|\theta)}; \tag{11.11}$$

its estimate $\hat{x}$ is given by an obvious analogue of (11.10). The advantage of (11.11) relative to (11.8) is that the dimensionality of θ_t is often much smaller than that of α_t. In the important case in which y_t is univariate, θ_t is a scalar. Furthermore, once a sample is available for θ, we can deduce a sample from α using the argument in Subsection 4.13.6.

11.3 Choice of an importance density

For the computation of the estimate (11.10), we need to sample N time series for $\alpha_1, \ldots, \alpha_n$ from the importance density $g(\alpha|Y_n)$. We need to choose this density carefully. To obtain a feasible procedure, sampling from $g(\alpha|Y_n)$ should be relatively easy and computationally fast. The importance density $g(\alpha|Y_n)$ should also be sufficiently close to $p(\alpha|Y_n)$ such that the Monte Carlo variance of $\hat{x}$ is small. Hence the choice of the importance density is essential for the quality of the estimate (11.10). For example, consider the choice of $g(\alpha|Y_n)$ being equal to $p(\alpha)$. The simulation from $p(\alpha)$ is simple by all means since we can rely directly on the model specification for α_t in (11.1) or (11.2). All selected α's from this density will however have no relation with the observed vector Y_n. Almost all draws will have no support from $p(\alpha, Y_n)$ and we will obtain a poor estimate of $x(\alpha)$ with high Monte Carlo variance. A more succesful choice is to consider models for the observation density $g(Y_n|\alpha)$ and the state density $g(\alpha)$ that are both linear Gaussian; it implies that the importance density $g(\alpha|Y_n) = g(Y_n|\alpha)\, g(\alpha) \,/\, g(Y_n)$ is Gaussian also, where $g(Y_n)$ is the likelihood function that is not relevant for α. An importance density should be chosen within the class of linear Gaussian state space models that closely resembles or approximates the model density $p(\alpha, Y_n)$. From the discussion in Section 4.9, we have learned that simulating from $g(\alpha|Y_n)$ is feasible by means of a simulation smoothing algorithm.

In Sections 10.6 and 10.7 it is shown how the mode of the smoothed density $p(\alpha|Y_n)$ can be obtained. The mode is obtained by repeated linearisations of the smoothed density around a trial estimate of its mode for α. The linearisation is based on an approximating linear model and allows the use of the Kalman filter and smoother. After convergence to the mode, the approximating model can effectively be treated as the importance density $g(\alpha|Y_n)$. Given the linear model, we can apply a simulation smoother to generate importance samples from the density $g(\alpha|Y_n)$.

Alternative choices of an importance density for nonlinear non-Gaussian state space models are proposed by Danielsson and Richard (1993), Liesenfeld and Richard (2003) and Richard and Zhang (2007); they refer to their method as *efficient importance sampling*. Their importance densities are based on the minimisation of the variance of the importance weights in logs, that is the variance of $\log w(\alpha, Y_n)$ where $w(\alpha, Y_n)$ is defined in (11.8). The construction of such an importance density requires simulation and is computationally demanding. Lee and Koopman (2004) compare the performances of the simulation-based methods when different importance densities are adopted. Koopman, Lucas and Scharth (2011) show that efficient importance sampling can also be implemented by using numerical integration, simulation smoothing and control variables. They show that their *numerically accelarated importance sampling* method leads to significant computational and numerical efficiency gains when compared to other importance sampling methods. We notice that a control variable is a traditional device for improving the efficiency of estimates obtained by simulation; see also the discussion in Durbin and Koopman (2000). A related device is the antithetic variable which we discuss in Subsection 11.4.3.

11.4 Implementation details of importance sampling

11.4.1 Introduction

In this section we describe the practical implementation details of importance sampling for our general class of models. The first step is to select an appropriate importance density $g(\alpha|Y_n)$ from which it is practical to generate samples of α or, when appropriate, θ. The next step is to express the relevant formulae in terms of variables which are as simple as possible; we do this in Subsection 11.4.2. In 11.4.3 we describe antithetic variables, which increase the efficiency of the simulation by introducing a balanced structure into the simulation sample. Questions of initialisation of the approximating linear Gaussian model are considered in Subsection 11.4.4. We can take a moderate value for N when computing $\hat{x}$ in practice; typical values are $N = 100$ and $N = 250$.

11.4.2 Practical implementation of importance sampling

Up to this point we have based our exposition of the ideas underlying the use of importance sampling on α and Y_n since these are the basic vectors of interest in the state space model. However, for practical computations it is important to express formulae in terms of variables that are as simple as possible. In particular, in place of the α_t's it is usually more convenient to work with the state disturbance terms $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$. We therefore consider how to reformulate the previous results in terms of η rather than α.

By repeated substitution from the relation $\alpha_{t+1} = T_t\alpha_t + R_t\eta_t$, for $t = 1, \dots, n$, we express $x(\alpha)$ as a function of α_1 and η; for notational convenience and because we intend to deal with the initialisation in Subsection 11.4.4, we suppress the dependence on α_1 and write $x(\alpha)$ as a function of η in the form $x^*(\eta)$. We next note that we could have written (11.3) in the form

$$\bar{x} = \mathrm{E}[x^*(\eta)|Y_n] = \int x^*(\eta)p(\eta|Y_n)\, d\eta. \tag{11.12}$$

Analogously to (11.8) we have

$$\bar{x} = \frac{\mathrm{E}_g[x^*(\eta)w^*(\eta, Y_n)]}{\mathrm{E}_g[w^*(\eta, Y_n)]}, \qquad w^*(\eta, Y_n) = \frac{p(\eta, Y_n)}{g(\eta, Y_n)}. \tag{11.13}$$

In this formula, E_g denotes expectation with respect to importance density $g(\eta|Y_n)$, which is the conditional density of η given Y_n in the approximating model, and

$$p(\eta, Y_n) = \prod_{t=1}^{n} p(\eta_t)p(y_t|\theta_t),$$

where $\theta_t = Z_t \alpha_t$. In the special case where $y_t = \theta_t + \varepsilon_t$, $p(y_t|\theta_t) = p(\varepsilon_t)$. In a similar way, for the same special case,

$$g(\eta, Y_n) = \prod_{t=1}^{n} g(\eta_t) g(\varepsilon_t).$$

For cases where the state equation is not linear and Gaussian, formula (11.13) provides the basis for the simulation estimates. When the state is linear and Gaussian, $p(\eta_t) = g(\eta_t)$ so in place of $w^*(\eta, Y_n)$ in (11.13) we take

$$w^*(\theta, Y_n) = \prod_{t=1}^{n} \frac{p(y_t|\theta_t)}{g(\varepsilon_t)}. \tag{11.14}$$

For the case $p(\eta_t) = g(\eta_t)$ and $y_t = \theta_t + \varepsilon_t$, we replace $w^*(\eta, Y_n)$ by

$$w^*(\varepsilon) = \prod_{t=1}^{n} \frac{p(\varepsilon_t)}{g(\varepsilon_t)}. \tag{11.15}$$

11.4.3 Antithetic variables

The simulations are based on random draws of η from the importance density $g(\eta|Y_n)$ using the simulation smoother as described in Section 4.9; this computes efficiently a draw of η as a linear function of rn independent standard normal deviates where r is the dimension of vector η_t and n is the number of observations. Efficiency is increased by the use of antithetic variables. An *antithetic variable* in this context is a function of a random draw of η which is equiprobable with η and which, when included together with η in the estimate of $\bar{x}$ increases the efficiency of the estimation. We assume that the importance density is Gaussian. We shall employ two types of antithetic variables. The first is the standard one given by $\check{\eta} = 2\hat{\eta} - \eta$ where $\hat{\eta} = \mathrm{E}_g(\eta|Y_n)$ is obtained from the disturbance smoother as described in Section 4.5. Since $\check{\eta} - \hat{\eta} = -(\eta - \hat{\eta})$ and η is normal, the two vectors η and $\check{\eta}$ are equi-probable. Thus we obtain two simulation samples from each draw of the simulation smoother; moreover, values of conditional means calculated from the two samples are negatively correlated, giving further efficiency gains. When this antithetic is used we say that the simulation sample is *balanced for location*.

The second antithetic variable was developed by Durbin and Koopman (1997). Let u be the vector of rn $N(0,1)$ variables that is used in the simulation smoother to generate η and let $c = u'u$; then $c \sim \chi^2_{rn}$. For a given value of c let $q = \Pr(\chi^2_{rn} < c) = F(c)$ and let $\acute{c} = F^{-1}(1-q)$. Then as c varies, c and $\acute{c}$ have the same distribution. Now take, $\acute{\eta} = \hat{\eta} + \sqrt{\acute{c}/c}(\eta - \hat{\eta})$. Then $\acute{\eta}$ has the same distribution as η. This follows because c and $(\eta - \hat{\eta})/\sqrt{c}$ are independently distributed. Finally, take $\grave{\eta} = \hat{\eta} + \sqrt{\acute{c}/c}(\check{\eta} - \hat{\eta})$. When this antithetic is used we say that the simulation sample is *balanced for scale*. By using both antithetics we obtain a set of four equiprobable values of η for each run of the simulation smoother giving a simulation sample which is balanced for location and scale.

The number of antithetics can be increased without difficulty. For example, take c and q as above. Then q is uniformly distributed on $(0, 1)$ and we write $q \sim U(0, 1)$. Let $q_1 = q + 0.5$ modulo 1; then $q_1 \sim U(0, 1)$ and we have a balanced set of four $U(0, 1)$ variables, q, q_1, $1 - q$ and $1 - q_1$. Take $\acute{c} = F^{-1}(1 - q)$ as before and similarly $c_1 = F^{-1}(q_1)$ and $\acute{c}_1 = F^{-1}(1 - q_1)$. Then each of c_1 and $\acute{c}_1$ can be combined with η and $\breve{\eta}$ as was $\acute{c}$ previously and we emerge with a balanced set of eight equiprobable values of η for each simulation. In principle this process could be extended indefinitely by taking $q_1 = q$ and $q_{j+1} = q_j + 2^{-k}$ modulo 1, for $j = 1, \ldots, 2^{k-1}$ and $k = 2, 3, \ldots$; however, two or four values of q are probably enough in practice. By using the standard normal distribution function applied to elements of u, the same idea could be used to obtain a new balanced value η_1 from η so by taking $\breve{\eta}_1 = 2\hat{\eta} - \eta_1$ we would have four values of η to combine with the four values of c. In the following we will assume that we have generated N draws of η using the simulation smoother and the antithetic variables; this means that N is a multiple of the number of different values of η obtained from a single draw of the simulation smoother. For example, when 250 simulation samples are drawn by the smoother and the one or two basic antithetics are employed, one for location and the other for scale, $N = 1000$. In practice, we have found that satisfactory results are obtained by only using the two basic antithetics.

In theory, importance sampling could give an inaccurate result on a particular occasion if in the basic formulae (11.13) very high values of $w^*(\eta, Y_n)$ are associated with very small values of the importance density $g(\eta|Y_n)$ in such a way that together they make a significant contribution to $\bar{x}$, and if also, on this particular occasion, these values happen to be over- or under-represented; for further discussion of this point see Gelman, Carlin, Stern and Rubin (1995, p. 307). In practice, we have not experienced difficulties from this source in any of the examples we have considered.

11.4.4 Diffuse initialisation

We now consider the situation where the model is non-Gaussian and some elements of the initial state vector are diffuse, the remaining elements having a known joint density; for example, they could come from stationary series. Assume that α_1 is given by (5.2) with $\eta_0 \sim p_0(\eta_0)$ where $p_0(\cdot)$ is a known density. It is legitimate to assume that δ is normally distributed as in (5.3) since we intend to let $\kappa \to \infty$. The joint density of α and Y_n is

$$p(\alpha, Y_n) = p(\eta_0) g(\delta) \prod_{t=1}^{n} p(\eta_t) p(y_t|\theta_t), \qquad (11.16)$$

with $\eta_0 = R_0'(\alpha_1 - a), \delta = A'(\alpha_1 - a)$ and $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$ for $t = 1, \ldots, n$, since $p(y_t|\alpha_t) = p(y_t|\theta_t)$.

As in Sections 10.6 and 10.7, we can find the mode of $p(\alpha|Y_n)$ by differentiating $\log p(\alpha, Y_n)$ with respect to $\alpha_1, \ldots, \alpha_{n+1}$. For given κ the contribution from $\partial \log g(\delta)/\partial\alpha_1$ is $-A\delta/\kappa$ which $\to 0$ as $\kappa \to \infty$. Thus in the limit the mode equation is the same as (10.54) except that $\partial \log p(\alpha_1)/\partial\alpha_1$ is replaced by $\partial \log p(\eta_0)/\partial\alpha_1$. In the case that α_1 is entirely diffuse, the term $p(\eta_0)$ does not enter into (11.16), so the procedure given in Subsection 10.6.3 for finding the mode applies without change.

When $p(\eta_0)$ exists but is non-Gaussian, it is preferable to incorporate a normal approximation to it, $g(\eta_0)$ say, in the approximating Gaussian density $g(\alpha, Y_n)$, rather than include a linearised form of its derivative $\partial \log p(\eta_0)/\partial\eta_0$ within the linearisation of $\partial \log p(\alpha, Y_n)/\partial\alpha$. The reason is that we are then able to initialise the Kalman filter for the linear Gaussian approximating model by means of the standard initialisation routines developed in Chapter 5. For $g(\eta_0)$ we could take either the normal distribution with mean vector and variance matrix equal to those of $p(\eta_0)$ or with mean vector equal to the mode of $p(\eta_0)$ and variance matrix equal to $[-\partial^2 p(\eta_0)/\partial\eta_0\partial\eta_0']^{-1}$. For substitution in the basic formula (11.8) we take

$$w(\alpha, Y_n) = \frac{p(\eta_0)p(\alpha_2, \ldots, \alpha_{n+1}, Y_n|\eta_0)}{g(\eta_0)g(\alpha_2, \ldots, \alpha_{n+1}, Y_n|\eta_0)}, \tag{11.17}$$

since the denisties $p(\delta)$ and $g(\delta)$ are the same and therefore cancel out; thus $w(\alpha, Y_n)$ remains unchanged as $\kappa \to \infty$. The corresponding equation for (11.13) becomes simply

$$w^*(\eta, Y_n) = \frac{p(\eta_0)p(\eta_1, \ldots, \eta_n, Y_n)}{g(\eta_0)g(\eta_1, \ldots, \eta_n, Y_n)}. \tag{11.18}$$

While the expressions (11.17) and (11.18) are technically manageable, the practical worker may well believe in a particular situation that knowledge of $p(\eta_0)$ contributes such a small amount of information to the investigation that it can be simply ignored. In that event the factor $p(\eta_0)/g(\eta_0)$ disappears from (11.17), which amounts to treating the whole vector α_1 as diffuse, and this simplifies the analysis significantly. Expression (11.17) then reduces to

$$w(\alpha, Y_n) = \prod_{t=1}^{n} \frac{p(\alpha_t|\alpha_{t-1})p(y_t|\alpha_t)}{g(\alpha_t|\alpha_{t-1})g(y_t|\alpha_t)}.$$

Expression (11.18) reduces to

$$w^*(\eta, Y_n) = \prod_{t=1}^{n} \frac{p(\eta_t)p(y_t|\theta_t)}{g(\eta_t)g(y_t|\theta_t)},$$

with $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$ for $t = 1, \ldots, n$.

For nonlinear models, the initialisation of the Kalman filter is similar and the details are handled in the same way.

11.5 Estimating functions of the state vector

We will discuss the estimation of general functions of the state vector based on importance sampling for analysing data from non-Gaussian and nonlinear models. We start by showing how the method enables us to estimate mean and variance functions of the state vector using simulation and antithetic variables. We also derive estimates of the additional variances of estimates due to simulation. We use these results to obtain estimates of conditional densities and distribution functions of scalar functions of the state. Then we continue by investigating how the methods can be used for forecasting and estimating missing observations in a data set.

11.5.1 Estimating mean functions

We will consider details of the estimation of conditional means $\bar{x}$ of functions $x^*(\eta)$ of the stacked state error vector and the estimation of error variances of our estimates. Let

$$w^*(\eta) = \frac{p(\eta, Y_n)}{g(\eta, Y_n)},$$

taking the dependence of $w^*(\eta)$ on Y_n as implicit since Y_n is constant from now on. Then (11.13) gives

$$\bar{x} = \frac{\mathrm{E}_g\left[x^*(\eta) w^*(\eta)\right]}{\mathrm{E}_g\left[w^*(\eta)\right]}, \tag{11.19}$$

which is estimated by

$$\hat{x} = \frac{\sum_{i=1}^N x_i w_i}{\sum_{i=1}^N w_i}, \tag{11.20}$$

where

$$x_i = x^*\big(\eta^{(i)}\big), \qquad w_i = w^*\big(\eta^{(i)}\big) = \frac{p\big(\eta^{(i)}, Y_n\big)}{g\big(\eta^{(i)}, Y_n\big)},$$

and $\eta^{(i)}$ is the ith draw of η from the importance density $g(\eta|Y_n)$ for $i = 1, \ldots, N$.

11.5.2 Estimating variance functions

For the case where $x^*(\eta)$ is a vector we could at this point present formulae for estimating the matrix $\mathrm{Var}[x^*(\eta)|Y_n]$ and also the variance matrix due to simulation of $\hat{x} - \bar{x}$. However, from a practical point of view the covariance terms are of little interest so it seems sensible to focus on variance terms by taking $x^*(\eta)$ as a scalar for estimation of variances; extension to include covariance terms is straightforward. We estimate $\mathrm{Var}[x^*(\eta)|Y_n]$ by

$$\widehat{\mathrm{Var}}[x^*(\eta)|Y_n] = \frac{\sum_{i=1}^N x_i^2 w_i}{\sum_{i=1}^N w_i} - \hat{x}^2. \tag{11.21}$$

The estimation error due to the simulation is

$$\hat{x} - \bar{x} = \frac{\sum_{i=1}^{N} w_i(x_i - \bar{x})}{\sum_{i=1}^{N} w_i}.$$

To estimate the variance of this, consider the introduction of the antithetic variables as described in Subsection 11.4.3 and for simplicity restrict the exposition to the case of the two basic antithetics for location and scale; the extension to a larger number of antithetics is straightforward. Denote the sum of the four values of $w_i(x_i - \bar{x})$ that come from the jth run of the simulation smoother by v_j and the sum of the corresponding values of $w_i(x_i - \hat{x})$ by $\hat{v}_j$. For N large enough, since the draws from the simulation smoother are independent, the variance due to simulation is, to a good approximation,

$$\mathrm{Var}_s(\hat{x}) = \frac{1}{4N} \frac{\mathrm{Var}(v_j)}{[\mathrm{E}_g\{w^*(\eta)\}^2]}, \tag{11.22}$$

which we estimate by

$$\widehat{\mathrm{Var}}_s(\hat{x}) = \frac{\sum_{j=1}^{N/4} \hat{v}_j^2}{\left(\sum_{i=1}^{N} w_i\right)^2}. \tag{11.23}$$

The ability to estimate simulation variances so easily is an attractive feature of our methods.

11.5.3 Estimating conditional densities

When $x^*(\eta)$ is a scalar function the above technique can be used to estimate the conditional distribution function and the conditional density function of x given Y_n. Let $G[x|Y_n] = \Pr[x^*(\eta) \leq x|Y_n]$ and let $I_x(\eta)$ be an indicator which is unity if $x^*(\eta) \leq x$ and is zero if $x^*(\eta) > x$. Then $G(x|Y_n) = \mathrm{E}_g(I_x(\eta)|Y_n)$. Since $I_x(\eta)$ is a function of η we can treat it in the same way as $x^*(\eta)$. Let S_x be the sum of the values of w_i for which $x_i \leq x$, for $i = 1, \dots, N$. Then estimate $G(x|Y_n)$ by

$$\hat{G}(x|Y_n) = \frac{S_x}{\sum_{i=1}^{N} w_i}. \tag{11.24}$$

This can be used to estimate quantiles. We order the values of x_i and we order the corresponding values w_i accordingly. The ordered sequences for x_i and w_i are denoted by $x_{[i]}$ and $w_{[i]}$, respectively. The $100k$% quantile is given by $x_{[m]}$ which is chosen such that

$$\frac{\sum_{i=1}^{m} w_{[i]}}{\sum_{i=1}^{N} w_{[i]}} \approx k.$$

We may interpolate between the two closest values for m in this approximation to estimate the $100k$% quantile. The approximation error becomes smaller as N increases.

11.5.4 Estimating conditional distribution functions

Similarly, if δ is the interval $(x - \frac{1}{2}d, x + \frac{1}{2}d)$ where d is suitably small and positive, let S^δ be the sum of the values of w_i for which $x^*(\eta) \in \delta$. Then the estimate of the conditional density $p(x|Y_n)$ of x given Y_n is

$$\hat{p}(x|Y_n) = d^{-1} \frac{S^\delta}{\sum_{i=1}^N w_i}. \tag{11.25}$$

This estimate can be used to construct a histogram.

We now show how to generate a sample of M independent values from the estimated conditional distribution of $x^*(\eta)$ using importance resampling; for further details of the method see Gelfand and Smith (1999) and Gelman, Carlin, Stern and Rubin (1995). Take $x^{[k]} = x_j$ with probability $w_j / \sum_{i=1}^N w_i$ for $j = 1, \ldots, N$. Then

$$\Pr\big(x^{[k]} \leq x\big) = \frac{\sum_{x_j \leq x} w_j}{\sum_{i=1}^N w_i} = \hat{G}(x|Y_n).$$

Thus $x^{[k]}$ is a random draw from the distribution function given by (11.24). Doing this M times with replacement gives a sample of $M \leq N$ independent draws. The sampling can also be done without replacement but the values are not then independent.

11.5.5 Forecasting and estimating with missing observations

The treatment of missing observations and forecasting by the methods of this chapter is straightforward. For missing observations, our objective is to estimate $\bar{x} = \int x^*(\eta) p(\eta|Y_n)\, d\eta$ where the stacked vector Y_n contains only those observational elements actually observed. We achieve this by omitting from the linear Gaussian approximating model the observational components that correspond to the missing elements in the original model. Only the Kalman filter and smoother algorithms are needed in the determination of the approximating model and we described in Section 4.10 how the filter is modified when observational vectors or elements are missing. For the simulation, the simulation smoother of Section 4.9 must be similarly modified to allow for the missing elements.

For forecasting, our objective is to estimate $\bar{y}_{n+j} = \mathrm{E}(y_{n+j}|Y_n)$, $j = 1, \ldots, J$, where we assume that $y_{n+1}, \ldots y_{n+J}$ and $\alpha_{n+2}, \ldots, \alpha_{n+J}$ have been generated by model (9.3) and (9.4), noting that α_{n+1} has already been generated by (9.4) with $t = n$. It follows from (9.3) that

$$\bar{y}_{n+j} = \mathrm{E}[\mathrm{E}(y_{n+j}|\theta_{n+j})|Y_n], \tag{11.26}$$

for $j = 1, \ldots, J$, where $\theta_{n+j} = Z_{n+j}\alpha_{n+j}$, with $Z_{n+1}, \ldots, Z_{n+J}$ assumed known. We estimate this as in Section 11.5 with $x^*(\eta) = \mathrm{E}(y_{n+j}|\theta_{n+j})$, extending the simulation smoother for $t = n + 1, \ldots, n + J$.

For exponential families,

$$\mathrm{E}(y_{n+j}|\theta_{n+j}) = \dot{b}_{n+j}(\theta_{n+j}),$$

as in Section 9.3 for $t \leq n$, so we take $x^*(\eta) = \dot{b}_{n+j}(\theta_{n+j})$, for $j = 1, \ldots, J$. For the model $y_t = \theta_t + \varepsilon_t$ in (9.7) we take $x^*(\eta) = \theta_t$.

11.6 Estimating loglikelihood and parameters

In this section we consider the estimation of the parameter vector ψ by maximum likelihood. Since analytical methods are not feasible we employ techniques based on simulation using importance sampling. We shall find that the techniques we develop are closely related to those we employed earlier in this chapter for estimation of the mean of $x(\alpha)$ given Y_n. Monte Carlo estimation of ψ by maximum likelihood using importance sampling was considered briefly by Shephard and Pitt (1997) and in more detail by Durbin and Koopman (1997) for the special case where $p(y_t|\theta_t)$ is non-Gaussian but α_t is generated by a linear Gaussian model. In this section we will begin by considering first the general case where both $p(y_t|\theta_t)$ and the state error density $p(\eta_t)$ in (9.4) are non-Gaussian and will specialise later to the simpler case where $p(\eta_t)$ is Gaussian. We will also consider the case where the state space models are nonlinear. Our approach will be to estimate the loglikelihood by simulation and then to estimate ψ by maximising the resulting value numerically.

11.6.1 Estimation of likelihood

The likelihood $L(\psi)$ is defined by $L(\psi) = p(Y_n|\psi)$, where for convenience we suppress the dependence of $L(\psi)$ on Y_n, so we have

$$L(\psi) = \int p(\alpha, Y_n)\, d\alpha.$$

Dividing and multiplying by the importance density $g(\alpha|Y_n)$ as in Section 11.2 gives

$$\begin{aligned} L(\psi) &= \int \frac{p(\alpha, Y_n)}{g(\alpha|Y_n)} g(\alpha|Y_n)\, d\alpha \\ &= g(Y_n) \int \frac{p(\alpha, Y_n)}{g(\alpha, Y_n)} g(\alpha|Y_n)\, d\alpha \\ &= L_g(\psi)\, \mathrm{E}_g\left[w(\alpha, Y_n)\right], \end{aligned} \tag{11.27}$$

where $L_g(\psi) = g(Y_n)$ is the likelihood of the approximating linear Gaussian model that we employ to obtain the importance density $g(\alpha|Y_n)$, E_g denotes expectation with respect to density $g(\alpha|Y_n)$, and $w(\alpha, Y_n) = p(\alpha, Y_n)/g(\alpha, Y_n)$

as in (11.8). Indeed we observe that (11.27) is essentially equivalent to (11.6). We note the elegant feature of (11.27) that the non-Gaussian likelihood $L(\psi)$ has been obtained as an adjustment to the linear Gaussian likelihood $L_g(\psi)$, which is easily calculated by the Kalman filter; moreover, the adjustment factor $\mathrm{E}_g[w(\alpha, Y_n)]$ is readily estimable by simulation. Obviously, the closer the importance joint density $g(\alpha, Y_n)$ is to the non-Gaussian density $p(\alpha, Y_n)$, the smaller will be the simulation sample required.

For practical computations we follow the practice discussed in Subsection 11.4.2 and Section 11.5 of working with the signal $\theta_t = Z_t\alpha_t$ in the observation equation and the state disturbance η_t in the state equation, rather than with α_t directly, since these lead to simpler computational procedures. In place of (11.27) we therefore use the form

$$L(\psi) = L_g(\psi)\, \mathrm{E}_g[w^*(\eta, Y_n)], \tag{11.28}$$

where $L(\psi)$ and $L_g(\psi)$ are the same as in (11.27) but E_g and $w^*(\eta, Y_n)$ have the interpretations discussed in Subsection 11.4.2. We then suppress the dependence on Y_n and write $w^*(\eta)$ in place of $w^*(\eta, Y_n)$ as in Section 11.5. We employ antithetic variables as in Subsection 11.4.3, and analogously to (11.20) our estimate of $L(\psi)$ is

$$\hat{L}(\psi) = L_g(\psi)\bar{w}, \tag{11.29}$$

where $\bar{w} = (1/N)\sum_{i=1}^{N} w_i$, with $w_i = w^*(\eta^{(i)})$ where $\eta^{(1)}, \dots \eta^{(N)}$ is the simulation sample generated by the importance density $g(\eta|Y_n)$.

11.6.2 Maximisation of loglikelihood

We estimate ψ by the value $\hat{\psi}$ of ψ that maximises $\hat{L}(\psi)$. In practice, it is numerically more stable to maximise

$$\log \hat{L}(\psi) = \log L_g(\psi) + \log \bar{w}, \tag{11.30}$$

rather than to maximise $\hat{L}(\psi)$ directly because the likelihood value can become very large. Moreover, the value of ψ that maximises $\log \hat{L}(\psi)$ is the same as the value that maximises $\hat{L}(\psi)$.

To calculate $\hat{\psi}$, $\log \hat{L}(\psi)$ is maximised by any convenient iterative numerical optimisation technique, as discussed, for example in Subsection 7.3.2. To ensure stability of the iterative process, it is important to use the same random numbers from the simulation smoother for each value of ψ. To start the iteration, an initial value of ψ can be obtained by maximising the approximate loglikelihood

$$\log L(\psi) \approx \log L_g(\psi) + \log w(\hat{\eta}),$$

where $\hat{\eta}$ is the mode of $g(\eta|Y_n)$ that is determined during the process of approximating $p(\eta|Y_n)$ by $g(\eta|Y_n)$; alternatively, the more accurate non-simulated approximation given in expression (21) of Durbin and Koopman (1997) may be used.

11.6.3 Variance matrix of maximum likelihood estimate

Assuming that appropriate regularity conditions are satisfied, the estimate of the large-sample variance matrix of $\hat{\psi}$ is given by the standard formula

$$\hat{\Omega} = \left[-\frac{\partial^2 \log L(\psi)}{\partial\psi\partial\psi'} \right]^{-1} \Bigg|_{\psi = \hat{\psi}}, \tag{11.31}$$

where the derivatives of $\log L(\psi)$ are calculated numerically from values of ψ in the neighbourhood of $\hat{\psi}$.

11.6.4 Effect of errors in parameter estimation

In the above treatment we have performed classical analyses in the traditional way by first assuming that the parameter vector is known and then substituting the maximum likelihood estimate $\hat{\psi}$ for ψ. The errors $\hat{\psi} - \psi$ give rise to biases in the estimates of functions of the state and disturbance vectors, but since the biases are of order n^{-1} they are usually small enough to be neglected. It may, however, be important to investigate the amount of bias in particular cases. In Subsection 7.3.7 we described techniques for estimating the bias for the case where the state space model is linear and Gaussian. Exactly the same procedure can be used for estimating the bias due to errors $\hat{\psi} - \psi$ for the non-Gaussian and nonlinear models considered in this chapter.

11.6.5 Mean square error matrix due to simulation

We have denoted the estimate of ψ that is obtained from the simulation by $\hat{\psi}$; let us denote by $\tilde{\psi}$ the 'true' maximum likelihood estimate of ψ that would be obtained by maximising the exact $\log L(\psi)$ without simulation, if this could be done. The error due to simulation is $\hat{\psi} - \tilde{\psi}$, so the mean square error matrix is

$$\text{MSE}(\hat{\psi}) = \text{E}_g[(\hat{\psi} - \tilde{\psi})(\hat{\psi} - \tilde{\psi})'].$$

Now $\hat{\psi}$ is the solution of the equation

$$\frac{\partial \log \hat{L}(\psi)}{\partial\psi} = 0,$$

which on expansion about $\tilde{\psi}$ gives approximately,

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial\psi} + \frac{\partial^2 \log \hat{L}(\tilde{\psi})}{\partial\psi\partial\psi'}(\hat{\psi} - \tilde{\psi}) = 0,$$

where

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial\psi} = \frac{\partial \log \hat{L}(\psi)}{\partial\psi}\Bigg|_{\psi=\tilde{\psi}}, \qquad \frac{\partial^2 \log \hat{L}(\tilde{\psi})}{\partial\psi\partial\psi'} = \frac{\partial^2 \log \hat{L}(\psi)}{\partial\psi\partial\psi'}\Bigg|_{\psi=\tilde{\psi}},$$

giving

$$\hat{\psi} - \tilde{\psi} = \left[-\frac{\partial^2 \log \hat{L}(\tilde{\psi})}{\partial \psi \partial \psi'} \right]^{-1} \frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi}.$$

Thus to a first approximation we have

$$\mathrm{MSE}(\hat{\psi}) = \hat{\Omega}\, \mathrm{E}_g \left[\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} \frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi'} \right] \hat{\Omega}, \tag{11.32}$$

where $\hat{\Omega}$ is given by (11.31).

From (11.30) we have

$$\log \hat{L}(\tilde{\psi}) = \log L_g(\tilde{\psi}) + \log \bar{w},$$

so

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} = \frac{\partial \log L_g(\tilde{\psi})}{\partial \psi} + \frac{1}{\bar{w}} \frac{\partial \bar{w}}{\partial \psi}.$$

Similarly, for the true loglikelihood $\log L(\tilde{\psi})$ we have

$$\frac{\partial \log L(\tilde{\psi})}{\partial \psi} = \frac{\partial \log L_g(\tilde{\psi})}{\partial \psi} + \frac{\partial \log \mu_w}{\partial \psi},$$

where $\mu_w = \mathrm{E}_g(\bar{w})$. Since $\tilde{\psi}$ is the 'true' maximum likelihood estimator of ψ,

$$\frac{\partial \log L(\tilde{\psi})}{\partial \psi} = 0.$$

Thus

$$\frac{\partial \log L_g(\tilde{\psi})}{\partial \psi} = -\frac{\partial \log \mu_w}{\partial \psi} = -\frac{1}{\mu_w} \frac{\partial \mu_w}{\partial \psi},$$

so we have

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} = \frac{1}{\bar{w}} \frac{\partial \bar{w}}{\partial \psi} - \frac{1}{\mu_w} \frac{\partial \mu_w}{\partial \psi}.$$

It follows that, to a first approximation,

$$\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} = \frac{1}{\bar{w}} \frac{\partial}{\partial \psi} (\bar{w} - \mu_w),$$

and hence

$$\mathrm{E}_g \left[\frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi} \frac{\partial \log \hat{L}(\tilde{\psi})}{\partial \psi'} \right] = \frac{1}{\bar{w}^2} \mathrm{Var} \left(\frac{\partial \bar{w}}{\partial \psi} \right).$$

Taking the case of two antithetics, denote the sum of the four values of w obtained from each draw of the simulation smoother by w_j^* for $j = 1, \dots, N/4$. Then $\bar{w} = N^{-1}\sum_{j=1}^{N/4} w_j^*$, so

$$\mathrm{Var}\left(\frac{\partial \bar{w}}{\partial \psi}\right) = \frac{4}{N}\mathrm{Var}\left(\frac{\partial w_j^*}{\partial \psi}\right).$$

Let $q^{(j)} = \partial w_j^* / \partial \psi$, which we calculate numerically at $\psi = \hat{\psi}$, and let $\bar{q} = (4/N)\sum_{j=1}^{N/4} q^{(j)}$. Then estimate (11.32) by

$$\widehat{\mathrm{MSE}}(\hat{\psi}) = \hat{\Omega}\left[\left(\frac{4}{N\bar{w}}\right)^2 \sum_{j=1}^{N/4}\left(q^{(j)} - \bar{q}\right)\left(q^{(j)} - \bar{q}\right)'\right]\hat{\Omega}. \tag{11.33}$$

The square roots of the diagonal elements of (11.33) may be compared with the square roots of the diagonal elements of $\hat{\Omega}$ in (11.31) to obtain relative standard errors due to simulation.

11.7 Importance sampling weights and diagnostics

In this section we briefly draw attention to ways to validate the effectiveness of the importance sampling method. The importance weight function $w(\alpha, Y_n)$ as defined in (11.8) is instrumental for this purpose. Geweke (1989) argued that importance sampling should only be used in settings where the variance of the importance weights is known to exist. Failure of this condition can lead to slow and unstable convergence of the estimator as the central limit theorem governing convergence fails to hold. Robert and Casella (2010, §4.3) provide examples of importance samplers that fail this condition and show that ignoring the problem can result in strongly biased estimators. While the variance conditions can be checked analytically in low dimensional problems, proving that they are met in high dimensional cases such as time series can be challenging.

Monahan (1993, 2001) and Koopman, Shephard and Creal (2009) have developed diagnostic procedures to check for the existence of the variance of the importance weights. It is based on the application of extreme value theory. Limit results from extreme value theory imply that we can learn about the variance of the importance weights by studying the behaviour of their distribution in the right hand tail. Test statistics are then formulated to test whether the tail of the distribution allows for a properly defined variance. If the characteristics of the tails are not sufficient, we reject the hypothesis that the variance of the importance weights exists. A set of graphical diagnostics can be deducted from the hypothesis and a complete insight can be obtained.

12 Particle filtering

12.1 Introduction

In this chapter we discuss the filtering of non-Gaussian and nonlinear series by fixing the sample at the values previously obtained at times $\ldots, t-2, t-1$ and choosing a fresh value at time t only. A new recursion over time is then required for the resulting simulation. The method is called *particle filtering*. We derive the results by classical analysis but point out that analogous methods can be obtained by linear methods and Bayesian analysis using Lemmas 2, 3 and 4 of Chapter 4. Illustrations are applied to real data in Chapter 14.

A large body of literature on particle filtering has emerged since the 1990s. An early use of the idea was given by Gordon, Salmond and Smith (1993) while the term particle appears to have first been used in this connection by Kitagawa (1996). For reviews of this field and resources of references, we refer to the book of Doucet, De Freitas and Gordon (2001) and also to the overview articles of Arulampalam, Maskell, Gordon and Clapp (2002), Maskell (2004) and Creal (2012).

We first consider in Section 12.2 filtering by the method of Chapter 11. We choose a sample from $y_1, \ldots, y_t$ and use importance sampling to estimate $x_{t+1}, x_{t+2}, \ldots$. This is a valid method of filtering and it is occasionally useful so we describe it here. However, it involves much more computing than particle filtering for routine use. We therefore go on to consider particle filtering and its association with importance sampling.

In Section 12.3 we discuss resampling techniques designed to reduce degeneracy in sampling. We go on in Section 12.4 to describe six methods of particle filtering, namely bootstrap filtering, auxiliary particle filtering, the extended particle filter, the unscented particle filter, the local regression filter and the mode equalisation filter.

12.2 Filtering by importance sampling

We shall consider a variety of formulations of the non-Gaussian nonlinear model, the most basic having the form

$$
\begin{aligned}
& p[y_t|Z_t(\alpha_t), \varepsilon_t], \qquad && \varepsilon_t \sim p(\varepsilon_t), && (12.1)\\
\alpha_{t+1} &= T_t(\alpha_t) + R_t(\alpha_t)\eta_t, \qquad && \eta_t \sim p(\eta_t), && (12.2)
\end{aligned}
$$

for $t = 1, 2, \ldots$, where Z_t, T_t and R_t are known matrix functions of α_t and where ε_t and η_t are disturbance series; as before, y_t and α_t are the observation and state series. We adopt the same notation as in the previous chapters. For convenience we define a collection of state vectors by

$$\alpha_{1:t} = (\alpha_1', \ldots, \alpha_t')', \tag{12.3}$$

whereas we keep the notation $Y_t = (y_1', \ldots, y_t')'$.

In this section we discuss the use of importance sampling for the filtering of non-Gaussian and nonlinear time series which are generated by the model (12.1) and (12.2). Consider an arbitrary function $x_t(\alpha_{1:t})$ of $\alpha_{1:t}$ in (12.3); as for the smoothing case introduced in Section 11.2, most of the problems we shall consider, from both classical and Bayesian perspectives, amount essentially to the estimation of the conditional mean

$$\begin{aligned} \bar{x}_t &= \mathrm{E}[x_t(\alpha_{1:t})|Y_t] \\ &= \int x_t(\alpha_{1:t}) p(\alpha_{1:t}|Y_t) \mathrm{d}\alpha_{1:t}, \end{aligned} \tag{12.4}$$

for $t = \tau + 1, \tau + 2, \ldots$ where τ is fixed and can be zero. We shall develop recursions for computing estimates of $\bar{x}_t$ by simulation; the sample $y_1, \ldots, y_\tau$ can possibly be used as a 'start-up' sample for initialising these recursions. We consider the estimation of $\bar{x}_t$ in (12.4) by simulation based on importance sampling analogously to the estimation of the conditional smoother $\bar{x}_t$ in (11.3) of Chapter 11. We let $g(\alpha_{1:t}|Y_t)$ be an importance density that is as close as possible to $p(\alpha_{1:t}|Y_t)$ subject to the requirement that sampling from $g(\alpha_{1:t}|Y_t)$ is sufficiently practical and inexpensive.

From (12.4) we have

$$\begin{aligned} \bar{x}_t &= \int x_t(\alpha_{1:t}) \frac{p(\alpha_{1:t}|Y_t)}{g(\alpha_{1:t}|Y_t)} g(\alpha_{1:t}|Y_t) \mathrm{d}\alpha_{1:t} \\ &= \mathrm{E}_g \left[x_t(\alpha_{1:t}) \frac{p(\alpha_{1:t}|Y_t)}{g(\alpha_{1:t}|Y_t)} \right], \end{aligned} \tag{12.5}$$

where E_g denotes expectation with respect to density $g(\alpha_{1:t}|Y_t)$. Since

$$p(\alpha_{1:t}, Y_t) = p(Y_t) p(\alpha_{1:t}|Y_t),$$

we obtain

$$\bar{x}_t = \frac{1}{p(Y_t)} \mathrm{E}_g \left[x_t(\alpha_{1:t}) \tilde{w}_t \right], \tag{12.6}$$

where

$$\tilde{w}_t = \frac{p(\alpha_{1:t}, Y_t)}{g(\alpha_{1:t}|Y_t)}. \tag{12.7}$$

For notational convenience we suppress the dependence of $\tilde{w}_t$ on $\alpha_{1:t}$ and Y_t in (12.6) and later. By putting $x_t(\alpha_{1:t}) = 1$, it follows from (12.6) that $p(Y_t) = \mathrm{E}_g(\tilde{w}_t)$. As a result, (12.6) can be written as

$$\bar{x}_t = \frac{\mathrm{E}_g\left[x_t(\alpha_{1:t})\tilde{w}_t\right]}{\mathrm{E}_g(\tilde{w}_t)}. \tag{12.8}$$

We propose to estimate (12.6) by means of a random sample $\alpha_{1:t}^{(1)}, \ldots, \alpha_{1:t}^{(N)}$ drawn from $g(\alpha_{1:t}|Y_t)$. We take as our estimator

$$\begin{aligned}\hat{x}_t &= \frac{N^{-1}\sum_{i=1}^N x_t(\alpha_{1:t}^{(i)})\tilde{w}_t^{(i)}}{N^{-1}\sum_{i=1}^N \tilde{w}_t^{(i)}} \\ &= \sum_{i=1}^N x_t(\alpha_{1:t}^{(i)})w_t^{(i)},\end{aligned} \tag{12.9}$$

where

$$\tilde{w}_t^{(i)} = \frac{p(\alpha_{1:t}^{(i)}, Y_t)}{g(\alpha_{1:t}^{(i)}|Y_t)}, \qquad w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^N \tilde{w}_t^{(j)}}. \tag{12.10}$$

The values $\tilde{w}_t^{(i)}$ are called *importance weights* and the values $w_t^{(i)}$ are called *normalised importance weights.* This treatment closely follows the basic ideas of importance sampling as discussed in Section 11.2. A simple method of filtering based on importance sampling is to draw fresh random samples of $\alpha_{1:t}^{(i)}$ from a suitable $g(\alpha_{1:t}|Y_t)$ at each time point t and estimate $\bar{x}_t$ by (12.9); however, for long time series this would be unduly laborious.

12.3 Sequential importance sampling

12.3.1 Introduction

To circumvent the simple method of filtering, it seems more natural in the context of filtering to retain the previous selection of $\alpha_{1:t-1}^{(i)}$ for each i and to confine the new sampling at time t to the selection of $\alpha_t^{(i)}$ only. We call this sequential process for choosing $\alpha_{1:t}^{(i)}$ and the estimation based on it *particle filtering*; the resulting sets of values $\alpha_{1:t}^{(1)}, \ldots, \alpha_{1:t}^{(N)}$ are called *particles*; thus the ith particle at time t is defined by the relation

$$\alpha_{1:t}^{(i)} = \left(\alpha_{1:t-1}^{(i)\prime}, \alpha_t^{(i)\prime}\right)',$$

where $\alpha_{1:t-1}^{(i)}$ is the ith particle at time $t-1$. The key to a filtering method is the development of recursions for selecting $\alpha_t^{(i)}$ and for computing the corresponding importance weights $\tilde{w}_t^{(i)}$ at time t and for $i = 1, \dots, N$. The weight $\tilde{w}_t^{(i)}$ is a function of $\alpha_{1:t-1}^{(i)}$ and Y_{t-1} together with the current observation y_t and a newly chosen $\alpha_t^{(i)}$. A 'start-up' sample for initialising the recursions is not needed.

12.3.2 Recursions for particle filtering

A new selection of $\alpha_t^{(i)}$'s need to be consistent with draws from the importance density $g(\alpha_{1:t}^{(i)}|Y_t)$. To construct a recursion for the importance density, let us temporarily drop the i index and denote a particular value of $\alpha_{1:t}^{(i)}$ by $\alpha_{1:t}$. We have

$$\begin{aligned} g(\alpha_{1:t}|Y_t) &= \frac{g(\alpha_{1:t}, Y_t)}{g(Y_t)} \\ &= \frac{g(\alpha_t|\alpha_{1:t-1}, Y_t)g(\alpha_{1:t-1}, Y_t)}{g(Y_t)} \\ &= g(\alpha_t|\alpha_{1:t-1}, Y_t)g(\alpha_{1:t-1}|Y_t). \end{aligned} \tag{12.11}$$

Now suppose that $\alpha_{1:t-1}$ is selected using knowledge only of Y_{t-1}. Moreover, given the realised values of $\alpha_{1:t-1}$ and Y_{t-1}, the value of the observational vector y_t has already been selected by a process which does not depend on the simulated sequence $\alpha_{1:t-1}$. Under these circumstances, the density $g(\alpha_{1:t-1}|Y_{t-1})$ is not affected by including y_t in its set of conditional variables Y_{t-1}. Hence, $g(\alpha_{1:t-1}|Y_t) \equiv g(\alpha_{1:t-1}|Y_{t-1})$. By adopting this equality and reversing the order in (12.11), we obtain

$$g(\alpha_{1:t}|Y_t) = g(\alpha_{1:t-1}|Y_{t-1})g(\alpha_t|\alpha_{1:t-1}, Y_t). \tag{12.12}$$

This is the fundamental recursion which underlines the practicality of particle filtering; see, for example, Doucet, De Freitas and Gordon (2001, §1.3). It is assumed that selecting $\alpha_t^{(i)}$ from the importance density $g(\alpha_t|\alpha_{1:t-1}, Y_t)$ is practical and inexpensive. In Section 12.4 we discuss different ways to sample from such an importance density.

We now develop a recursion for calculation of the weights $w_t^{(i)}$ in (12.10) for particle filtering. From (12.7) and (12.12),

$$\begin{aligned} \tilde{w}_t &= \frac{p(\alpha_{1:t}, Y_t)}{g(\alpha_{1:t}|Y_t)} \\ &= \frac{p(\alpha_{1:t-1}, Y_{t-1})p(\alpha_t, y_t|\alpha_{1:t-1}, Y_{t-1})}{g(\alpha_{1:t-1}|Y_{t-1})g(\alpha_t|\alpha_{1:t-1}, Y_t)}. \end{aligned}$$

Due to the Markovian nature of the model (12.1) and (12.2), we have

$$p(\alpha_t, y_t|\alpha_{1:t-1}, Y_{t-1}) = p(\alpha_t|\alpha_{t-1})p(y_t|\alpha_t),$$

so that

$$\tilde{w}_t = \tilde{w}_{t-1} \frac{p(\alpha_t|\alpha_{t-1})p(y_t|\alpha_t)}{g(\alpha_t|\alpha_{1:t-1}, Y_t)}.$$

For $\tilde{w}_t^{(i)}$ in (12.10) we therefore have the recursion

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-1}^{(i)} \frac{p(\alpha_t^{(i)}|\alpha_{t-1}^{(i)})p(y_t|\alpha_t^{(i)})}{g(\alpha_t^{(i)}|\alpha_{1:t-1}^{(i)}, Y_t)}, \tag{12.13}$$

from which we obtain the normalised weight $w_t^{(i)}$ by

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^N \tilde{w}_t^{(j)}}, \tag{12.14}$$

for $i = 1, \dots, N$ and $t = \tau + 1, \tau + 2, \dots$. The recursion (12.13) is initialised by $\tilde{w}_\tau^{(i)} = 1$ for $i = 1, \dots, N$. We then estimate x_t by $\hat{x}_t$ from (12.9) for which a value for $\alpha_t^{(i)}$ is selected from importance density $g(\alpha_t|\alpha_{1:t-1}, Y_t)$, for $i = 1, \dots, N$. The value $\tilde{w}_{t-1}^{(i)}$ in (12.13) can be replaced by $w_{t-1}^{(i)}$; this gives the same end result due to the normalisation. Importance sampling based on this approach is called *sequential importance sampling* (SIS). It was originally developed by Hammersley and Morton (1954) and applied to state space models by Handschin and Mayne (1969) and Handschin (1970).

12.3.3 Degeneracy and resampling

The following problem can occur in the practice of simulation when recursion (12.13) is adopted without modification. As t increases, the distribution of the weights $w_t^{(i)}$ becomes highly skewed. It is possible for all but one particle to have negligible weights, for t large. We then say that the sample has become *degenerate*. The problem of degeneracy typically occurs when the likelihood function $p(y_t|\alpha_t)$ is highly peaked relatively to the density $p(\alpha_t|\alpha_{t-1})$, with the effect that few of the values $\alpha_t^{(1)}, \dots, \alpha_t^{(N)}$ lead to non-neglible values of $w_t^{(i)}$. It is obviously wasteful to retain particles in the recursion that are contributing negligible weights to the estimate $\hat{x}_t$ in (12.9).

A way to combat this degeneracy is to proceed as follows. We first notice that $\hat{x}_t$ in (12.9) has the form of a weighted mean of a function $x_t(\alpha_{1:t}^{(i)})$ of a random vector $\alpha_{1:t}$ which takes values $\alpha_{1:t}^{(i)}$ with probabilities $w_t^{(i)}$ for $i = 1, \dots, N$ with $\sum_{i=1}^N w_t^{(i)} = 1$. Now take the values of $\tilde{\alpha}_{1:t}^{(i)} = \alpha_{1:t}^{(i)}$ that have a been selected as *old values*, and select *new values* $\alpha_{1:t}^{(i)}$ from the old values $\tilde{\alpha}_{1:t}^{(i)}$ with probabilities $w_t^{(1)}, \dots, w_t^{(N)}$ with replacement. This procedure is called *resampling*; it eliminates particles that have a negligible effect on the estimate.

The sample mean of the new $x_t(\alpha_{1:t}^{(i)})$'s is the old $\hat{x}_t$ defined by (12.9). In terms of the new $\alpha_{1:t}^{(i)}$'s we therefore may consider a new estimate of x_t,

$$\hat{x}_t = \frac{1}{N}\sum_{i=1}^{N} x_t(\alpha_{1:t}^{(i)}), \tag{12.15}$$

which has the same form as (12.9) with the normalised weights reset at $w_t^{(i)} = N^{-1}$. Although resampling is a technique that combats degeneracy in particle filtering, it clearly introduces additional Monte Carlo variation into the estimate $\hat{x}_t$ as computed by (12.15). It has been shown by Chopin (2004) that the estimate $\hat{x}_t$ computed before resampling as in (12.9) is more efficient and therefore is the preferred estimate. Resampling can take place after the computation of $\hat{x}_t$.

Where α_t is one-dimensional a gain in efficiency can be achieved by employing *stratified sampling* instead of random sampling for selection of the $\alpha_t^{(i)}$'s; this was pointed out by Kitagawa (1996). Here, we merely indicate the underlying idea. Let $W(\alpha)$ be the distribution function of the discrete distribution with probabilities $w_t^{(i)}$ at old values $\alpha_t^{(i)}$, that is $W(\alpha) = \sum_i w_t^{(i)}$ for all $\alpha_t^{(i)} \leq \alpha$. Choose a value c_1 from the uniform distribution between 0 and $1/N$. Then take as the new values $\alpha_t^{(i)}$ those values of α corresponding to $W(\alpha) = c_1 + (i-1)/N$ for $i = 1, \ldots, N$. Alternative resampling schemes are the systematic resampling method of Carpenter, Clifford and Fearnhead (1999) and the residual resampling method of Liu and Chen (1998). A detailed discussion on resampling with more relevant references to the statistical literature is given by Creal (2012).

Resampling can take place at each time t but it is not necessary. When sufficient particles remain into the next period, we can proceed without the resampling step. To assess whether a sufficient number of particles has remained, Liu and Chen (1998) introduce the *effective sample size* which is given by

$$ESS = \left(\sum_{i=1}^{N} w_t^{(i)\,2}\right)^{-1},$$

where $w_t^{(i)}$ is the normalised importance weight. By construction, ESS is a value between 1 and N and it measures weight stability. When only a few weights are relatively large, the ESS is small. When the weights are uniformly distributed, the value of ESS is close to N. In practice, it is often advocated that resampling should take place when $ESS < k \cdot N$ for some fraction $k = 0.75$ or $k = 0.5$.

While resampling increases the number of effective particles, it does not eliminate the degeneracy problem altogether. Consider, for example, the simple case where $x(\alpha_{1:t}) = \alpha_t$ and a particular value of (the old) $w_t^{(i)}$ is relatively high. Then in the resampling the corresponding value of $\alpha_t^{(i)}$ will be selected a large number of times so the variance of the estimate (12.15) is accordingly high. Nevertheless, it appears that on balance resampling is effective on a routine basis and we shall use it in all the examples considered in this chapter.

12.3.4 Algorithm for sequential importance sampling

For the implementation of sequential importance sampling, resampling does not take place with respect to the entire path $\alpha_{1:t}$ but only to the most recent value of α_t. In most cases of practical interest, the function $x_t(\alpha_{1:t})$ can be defined such that

$$x_t(\alpha_{1:t}) \equiv x_t(\alpha_t), \qquad t = 1, \ldots, n.$$

Furthermore, at time t, we are interested in estimating $x_t(\alpha_t)$ rather than $x_1(\alpha_1), \ldots, x_t(\alpha_t)$. Similarly, in the case of the Kalman filter, we also concentrate on the filtering distribution $p(\alpha_t|Y_t)$ rather than $p(\alpha_{1:t}|Y_t)$. It is shown by Chopin (2004) that the particle filter provides a consistent and asymptotically normal estimate of all moment characteristics of $p(\alpha_j|Y_t)$ only for $j = t$, not for $j < t$. In our particle filter implementations below, we therefore only resample α_t.

A formal description of the *sequential importance sampling resampling* (SISR) procedure is given by the following steps, for all $i = 1, \ldots, N$ and at fixed time t.

(i) Sample α_t : draw N values $\tilde{\alpha}_t^{(i)}$ from $g(\alpha_t|\alpha_{t-1}^{(i)}, Y_t)$ and store $\tilde{\alpha}_{t-1:t}^{(i)} = \left\{\alpha_{t-1}^{(i)}, \tilde{\alpha}_t^{(i)}\right\}$.

(ii) Weights : compute the corresponding weights $\tilde{w}_t^{(i)}$

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-1}^{(i)} \frac{p(\tilde{\alpha}_t^{(i)}|\alpha_{t-1}^{(i)})p(y_t|\tilde{\alpha}_t^{(i)})}{g(\tilde{\alpha}_t^{(i)}|\alpha_{t-1}^{(i)}, Y_t)}, \qquad i = 1, \ldots, N,$$

and normalise the weights to obtain $w_t^{(i)}$ as in (12.14).

(iii) Compute the variable of interest x_t: given the set of particles $\left\{\tilde{\alpha}_t^{(1)}, \ldots, \tilde{\alpha}_t^{(N)}\right\}$, compute

$$\hat{x}_t = \sum_{i=1}^{N} w_t^{(i)} x_t(\tilde{\alpha}_t^{(i)}).$$

(iv) Resample : draw N new independent particles $\alpha_t^{(i)}$ from $\left\{\tilde{\alpha}_t^{(1)}, \ldots, \tilde{\alpha}_t^{(N)}\right\}$ with replacement and with corresponding probabilities $\left\{w_t^{(1)}, \ldots, w_t^{(N)}\right\}$.

The procedure is recursively repeated for $t = \tau + 1, \tau + 2, \ldots, n$. The simulation from the importance density $g(\alpha_t|\alpha_{t-1}, Y_t)$ in step (i) is discussed in Section 12.4.

12.4 The bootstrap particle filter

12.4.1 Introduction

The selection of $\alpha_t^{(i)}$ given $\alpha_{1:t-1}^{(i)}$ and Y_{t-1} is crucial in particle filtering as it affects the computation of $\hat{x}_t$ and the recursive computation of the weights $\tilde{w}_t^{(i)}$, for $i = 1, \ldots, N$. In this section we present the main methodology of the particle filter based on a basic way of selecting the $\alpha_t^{(i)}$'s. The method is referrd to as the bootstrap filter. We shall focus attention on the development of the recursion (12.13), the estimation formula (12.15) and the computer algorithm for their implementations.

12.4.2 The bootstrap filter

The first particle filter to be developed was the *bootstrap filter* of Gordon, Salmond and Smith (1993). It is sometimes called the *sampling importance resampling* (SIR) filter; see, for example, Arulampalam, Maskell, Gordon and Clapp (2002). The key to the construction of a particle filter that relies on the weight recursion (12.13) is the choice of the importance density $g(\alpha_t|\alpha_{1:t-1}, Y_t)$. Since the joint density of α_t and y_t given $\alpha_{1:t-1}$ and Y_{t-1} depends only on α_{t-1}, we can restrict ourselves to importance densities of the form $g(\alpha_t|\alpha_{t-1}, y_t)$. Ideally we would take $g(\alpha_t|\alpha_{t-1}, y_t) = p(\alpha_t|\alpha_{t-1}, y_t)$ but this is normally not available in analytical form and so we look for approximations to it. The bootstrap filter takes the importance density as

$$g(\alpha_t|\alpha_{1:t-1}, y_t) = p(\alpha_t|\alpha_{t-1}). \tag{12.16}$$

At first sight this looks crude since it neglects relevant information in y_t but when used with resampling and with N large enough, it can work well in many cases of interest and it is widely used in practice. For the bootstrap filter, the recursion (12.13) therefore reduces to the simple form

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-1}^{(i)} p(y_t|\alpha_t^{(i)}).$$

We employ resampling at each time $t = 1, \ldots, n$. Since the weights are reset after the resampling of $\alpha_{t-1}^{(i)}$ at $w_{t-1}^{(i)} = 1/N$, the normalised weight becomes simply

$$w_t^{(i)} = \frac{p(y_t|\alpha_t^{(i)})}{\sum_{j=1}^{N} p(y_t|\alpha_t^{(j)})}, \qquad i = 1, \ldots, N. \tag{12.17}$$

The observation density $p(y_t|\alpha_t^{(i)})$ is evaluated straightforwardly for given $\alpha_t^{(i)}$.

12.4.3 Algorithm for bootstrap filter

Suppose that we have a *start up sample* $y_1, \ldots, y_\tau$ and that we intend to begin particle filtering at time $\tau + 1$. The algorithm for the recursive implementation of the filter then proceeds as follows, for all $i = 1, \ldots, N$ and at fixed time t.

(i) Sample α_t : draw N values $\tilde{\alpha}_t^{(i)}$ from $p(\alpha_t|\alpha_{t-1}^{(i)})$.

(ii) Weights : compute the corresponding weights $\tilde{w}_t^{(i)}$

$$\tilde{w}_t^{(i)} = p(y_t|\tilde{\alpha}_t^{(i)}), \qquad i = 1, \ldots, N,$$

and normalise the weights to obtain $w_t^{(i)}$ as in (12.14).

(iii) Compute the variable of interest x_t: given the set of particles $\left\{\tilde{\alpha}_t^{(1)}, \ldots, \tilde{\alpha}_t^{(N)}\right\}$, compute

$$\hat{x}_t = \sum_{i=1}^{N} w_t^{(i)} x_t(\tilde{\alpha}_t^{(i)}).$$

(iv) Resample : draw N new independent particles $\alpha_t^{(i)}$ from $\left\{\tilde{\alpha}_t^{(1)}, \ldots, \tilde{\alpha}_t^{(N)}\right\}$ with replacement and with corresponding probabilities $\left\{w_t^{(1)}, \ldots, w_t^{(N)}\right\}$.

We repeat these step for $t = \tau+1, \tau+2, \ldots$. The advantages of this filter are that it is quick and easy to operate and requires very little storage. The drawback is that if the likelihood $p(y_t|\alpha_t)$ is sharply peaked relative to the density $p(\alpha_t|\alpha_{t-1})$ there may be many repetitions of the likelier particles in step (iv), thus reducing the effective number of particles under consideration. A further weakness of the bootstrap filter arises from the fact that value of y_t is not taken into account in the selection of $\alpha_t^{(i)}$ in step (i) of the algorithm. Ways of taking y_t into account are proposed in the next subsections.

12.4.4 Illustration: local level model for Nile data

To illustrate the accuracy of the bootstrap filter for the local level model (2.3) and to show the importance of resampling, we have considered the Nile time series from Chapter 2. The local level model is $y_t = \alpha_t + \varepsilon_t$ with $\alpha_{t+1} = \alpha_t + \eta_t$ and $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ for $t = 1, \ldots, n$. The disturbances $\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ and $\eta_t \sim \mathrm{N}(0, \sigma_\eta^2)$ are mutually and serially independent. The filtered estimate of α_t using observations $y_1, \ldots, y_t$ together with its error variance, $a_{t|t}$ and $P_{t|t}$, respectively, are routinely computed by the Kalman filter (2.15). Alternatively, the filtered state estimate and its error variance can be computed by the bootstrap filter and we can compare its accuracy relative to the output of the Kalman filter.

We first implement the bootstrap filter without the resampling step (iv) and with $N = 10,000$. The variances σ_ε^2 and σ_η^2 of the local level model are set equal to their maximum likelihood estimates 15,099 and 1469.1 from Subsection 2.10.3, respectively. The remaining first three steps remain. The bootstrap filter for the local level model is given by

(i) Draw N values $\tilde{\alpha}_t^{(i)} \sim \mathrm{N}(\alpha_{t-1}^{(i)}, \sigma_\eta^2)$.

(ii) Compute the corresponding weights $\tilde{w}_t^{(i)}$

$$\tilde{w}_t^{(i)} = w_{t-1}^{(i)} \exp\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma_\varepsilon^2 - \frac{1}{2}\sigma_\varepsilon^{-2}(y_t - \tilde{\alpha}_t^{(i)})\right), \quad i = 1, \ldots, N,$$

and normalise the weights to obtain $w_t^{(i)}$ as in (12.14).

(iii) Compute

$$\hat{a}_{t|t} = \sum_{i=1}^{N} w_t^{(i)} \tilde{\alpha}_t^{(i)}, \qquad \hat{P}_{t|t} = \sum_{i=1}^{N} w_t^{(i)} \tilde{\alpha}_t^{(i)\,2} - \hat{a}_{t|t}^2.$$

(iv) Set $\alpha_t^{(i)} = \tilde{\alpha}_t^{(i)}$ for $i = 1, \ldots, N$ (without resampling).

In Fig. 12.1 we present the Nile data y_t, the filtered state estimate $\hat{a}_{t|t}$ and the corresponding confidence interval based on $\hat{P}_{t|t}$ for $t = 1, \ldots, n$. The panels (ii)

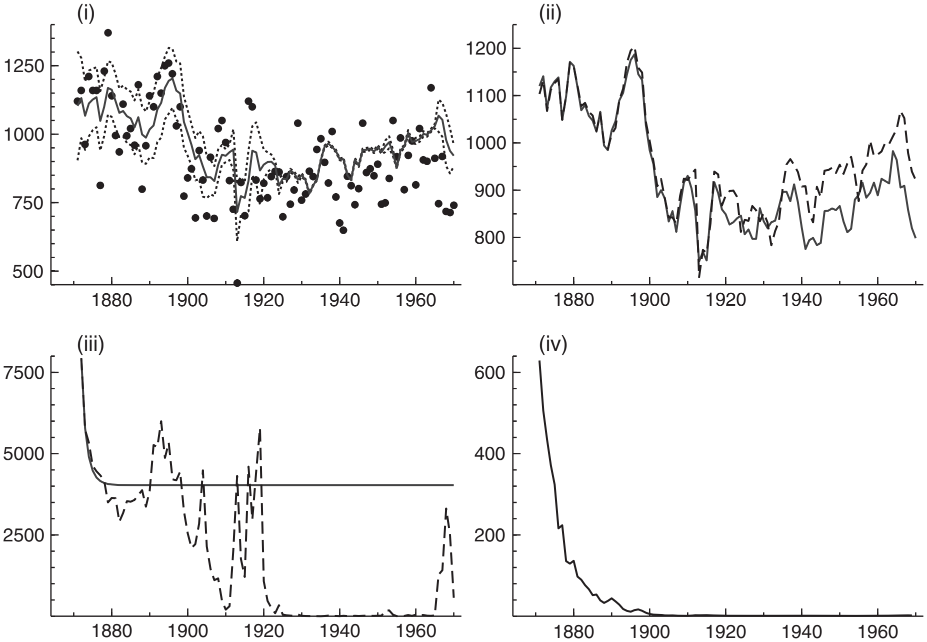

Fig. 12.1 Bootstrap filter with $N = 10,000$ **without** resampling for Nile data: (i) data (dots), filtered estimate $\hat{a}_{t|t}$ and its 90% confidence intervals; (ii) $\hat{a}_{t|t}$ (dashed) and $a_{t|t}$ (solid) from the Kalman filter; (iii) $\hat{P}_{t|t}$ and $P_{t|t}$; (iv) effective sample size ESS.

and (iii) present $a_{t|t}$ from the Kalman filter with $\hat{a}_{t|t}$ and $P_{t|t}$ from the Kalman filter with $\hat{P}_{t|t}$ for $t = 1, \dots, n$. It is clear that the differences are large and hence the bootstrap filter without resampling is inaccurate. The ESS time series in panel (iv) confirms this finding.

In Fig. 12.2 we present the same output as in Fig. 12.1 but now for the bootstrap filter with step (iv) replaced by the stratified resampling step proposed by Kitagawa (1996), that is

(iv) Select N new independent particles $\alpha_t^{(i)}$ using stratified sampling.

Panel (i) of Fig. 12.2 presents the filtered state estimate and the 90% confidence interval. From panel (ii) we learn that the filtered estimate $\hat{a}_{t|t}$ is virtually the same as $a_{t|t}$ from the Kalman filter. Also the variances in panel (iii) are very close to each other. We can conclude that resampling takes a crucial role in the bootstrap filter. The ESS time series in panel (iv) confirms this finding. The number of relevant particles remains high throughout the time series.

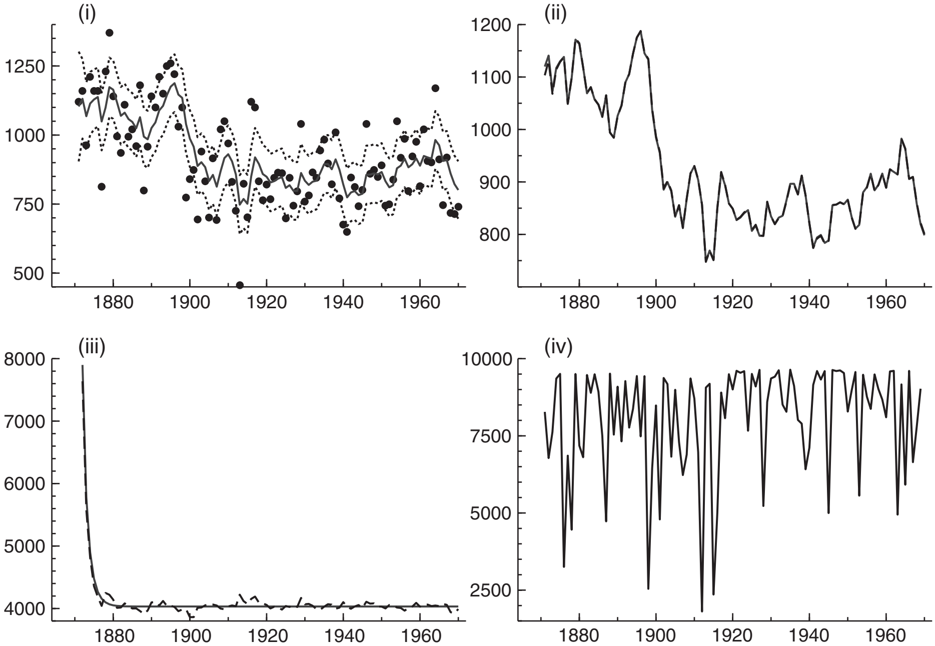

Fig. 12.2 Bootstrap filter with $N = 10{,}000$ **with** resampling for Nile data: (i) data (dots), filtered estimate $\hat{a}_{t|t}$ and its 90% confidence intervals; (ii) $\hat{a}_{t|t}$ (dashed) and $a_{t|t}$ (solid) from the Kalman filter; (iii) $\hat{P}_{t|t}$ and $P_{t|t}$; (iv) effective sample size ESS.

12.5 The auxiliary particle filter

The auxiliary particle filter was proposed by Pitt and Shephard (1999) as a development of the bootstrap filter. They gave a general treatment of the method whereas we will present only a simplified version below. The aim of their proposal is to reduce the number of repetitions in the final stage by introducing an additional selection stage involving the value of y_t. A modification of the auxiliary particle filter is considered by Johansen and Doucet (2008).

12.5.1 Algorithm for auxiliary filter

The primary modification of the auxiliary particle filter is based on finding an effective importance density at time t that also considers the knowledge of y_t for selecting new particles. We assume that the particles $\alpha_{1:t-1}^{(i)}$ are available at time $t-1$ with weights $w_{t-1}^{(i)} = 1/N$ for $i = 1, \ldots, N$. The auxiliary particle filter method proceeds as follows, for all $i = 1, \ldots, N$ and at fixed time t.

(i) Predict α_t: for each value $\alpha_{t-1}^{(i)}$, predict the corresponding value for α_t using a deterministic function (no sampling) and denote the prediction by $\alpha_t^{*(i)}$.

(ii) Intermediate weights: compute weights corresponding to prediction $\alpha_t^{*(i)}$,

$$\tilde{w}_t^{*(i)} = g(y_t|\alpha_t^{*(i)})w_{t-1}^{(i)},$$

and normalise to obtain $w_t^{*(i)}$.

(iii) Resample: draw N new particles $\tilde{\alpha}_{t-1}^{(i)}$, with replacement, from $\alpha_{t-1}^{(1)}, \ldots, \alpha_{t-1}^{(N)}$ with probabilities $w_t^{*(1)}, \ldots, w_t^{*(N)}$.

(iv) Sample α_t: draw N values $\alpha_t^{(i)}$ from $g(\alpha_t|\tilde{\alpha}_{t-1}^{(i)})$.

(v) Weights: compute the corresponding weights $\tilde{w}_t^{(i)}$

$$\tilde{w}_t^{(i)} = \frac{p(y_t|\alpha_t^{(i)})p(\alpha_t^{(i)}|\tilde{\alpha}_{t-1}^{(i)})}{g(y_t|\alpha_t^{*(i)})g(\alpha_t^{(i)}|\tilde{\alpha}_{t-1}^{(i)})}, \qquad i = 1, \ldots, N,$$

and normalise the weights to obtain $w_t^{(i)}$.

(vi) Compute the variable of interest x_t : given the set of particles $\left\{\alpha_t^{(1)}, \ldots, \alpha_t^{(N)}\right\}$, compute

$$\hat{x}_t = \sum_{i=1}^{N} w_t^{(i)} x_t(\alpha_t^{(i)}).$$

In step (i) we need to predict the state vector α_t given α_{t-1}. For the nonlinear state equation $\alpha_{t+1} = T_t(\alpha_t) + R_t(\eta_t)$, we can simply take $\alpha_t^{*(i)} = T_{t-1}(\alpha_{t-1}^{(i)})$ as our state prediction for $i = 1, \ldots, N$. Pitt and Shephard (1999) originally proposed the above initial prediction; Johansen and Doucet (2008) have considered other prediction choices and their statistical performances. The steps (iv), (v) and (vi) are essentially the same as the steps (i), (ii) and (iii) of the bootstrap filter algorithm, respectively.

Since the importance density $g(\alpha_t|\alpha_{t-1})$ is continuous there will, with probability one, be no omissions or repetitions in the values of $\alpha_t^{\sigma(i)}$ chosen in step (iv). Since the distribution of weights $w_t^{\sigma(i)}$ can be expected to be relatively lightly skewed, there should be fewer omissions and repetitions of values of $\alpha_t^{(i)}$ in the particles obtained from the auxiliary particle filter, as compared with those obtained from bootstrap filter. On the other hand, because of the way the $\alpha_{t-1}^{(i)}$'s are selected in step (iii), the distribution of the components $\alpha_{t-1}^{(i)}$ for the particles $\alpha_t^{(i)}$ should be comparable with those for the bootstrap filter.

12.5.2 Illustration: local level model for Nile data

To illustrate the possible gains in accuracy of the auxiliary particle filter in comparison with the bootstrap filter, we continue our illustration for the local level model (2.3) applied to the Nile time series from Chapter 2. The local level model is $y_t = \alpha_t + \varepsilon_t$ with $\alpha_{t+1} = \alpha_t + \eta_t$ and $\alpha_1 \sim \mathrm{N}(a_1, P_1)$ for $t = 1, \ldots, n$. The disturbances $\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ and $\eta_t \sim \mathrm{N}(0, \sigma_\eta^2)$ are mutually and serially independent. We aim to compute the filtered estimate of α_t using observations Y_t together with its error variance. The estimates of σ_ε^2 and σ_η^2 are obtained from the maximum likelihood method as reported in Subsection 2.10.3. The purpose is to compare the performances of the bootstrap and the auxiliary particle filters in computing the filtered estimates of α_t. In panels (i) and (iii) of Fig. 12.3 we present the output of the bootstrap and auxiliary filters, respectively. We can conclude that differences in the graphical output cannot be detected. In panels (ii) and (iv) the ESS for each time period of the bootstrap and auxiliary filters are presented, respectively. We have shown that the effective sample size of the auxiliary particle filter is higher for all time periods. In time periods where the ESS is relatively low for the bootstrap filter, the ESS for the auxiliary filter is at least twice as high as the ESS for the bootstrap filter.

12.6 Other implementations of particle filtering

We investigate other strategies for the selection of $\alpha_t^{(i)}$ given $\alpha_{1:t-1}^{(i)}$ and Y_{t-1} next.

12.6.1 Importance density from extended or unscented filter

The importance density $g(\alpha_t|\alpha_{t-1}, Y_t)$ can also be obtained from the extended Kalman filter or the unscented Kalman filter. Different ways of incorporating the ideas behind the extended or unscented filter on the one hand and particle

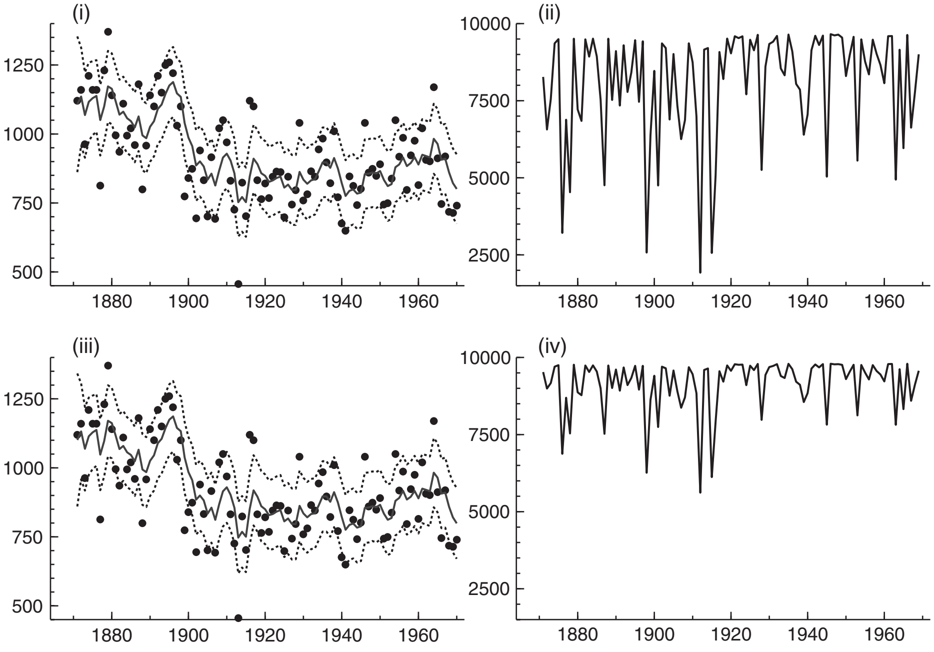

Fig. 12.3 Bootstrap filter versus auxiliary particle filter with $N = 10,000$ for Nile data: (i) data (dots), filtered estimate $\hat{a}_{t|t}$ and its 90% confidence intervals from bootstrap filter; (ii) effective sample size ESS for bootstrap filter; (iii) data (dots), filtered estimate $\hat{a}_{t|t}$ and its 90% confidence intervals from auxiliary filter; (iv) effective sample size ESS for auxiliary filter.

filtering on the other hand can be considered; see van der Merwe, Doucet and de Freitas (2000) for the first of such implementations.

In the treatment given below, we explore an alternative implementation in which we aim to incorporate y_t into the proposal density. Since the approximate filtering equations can be regarded as Taylor expansions of a certain order of $p(\alpha_t|\alpha_{t-1}, Y_t)$, we can expect such an importance density to be accurate. The extended or unscented filters can be introduced in different ways as we will discuss below.

The set of particles $\alpha_{t-1}^{(1)}, \ldots, \alpha_{t-1}^{(N)}$ provide information about the distribution of α_{t-1} given Y_{t-1} including its mean and variance, that is

$$\bar{a}^{+}_{t-1|t-1} = N^{-1} \sum_{i=1}^{N} \alpha_{t-1}^{(i)},$$

$$\bar{P}^{+}_{t-1|t-1} = N^{-1} \sum_{i=1}^{N} \left(\alpha_{t-1}^{(i)} - \bar{a}^{+}_{t-1|t-1} \right) \left(\alpha_{t-1}^{(i)} - \bar{a}^{+}_{t-1|t-1} \right)',$$

where $\bar{a}^+_{t-1|t-1}$ is the particle filter estimate of $\mathrm{E}(\alpha_{t-1}|Y_{t-1})$ and $\bar{P}^+_{t-1|t-1}$ is the variance estimate of $\mathrm{Var}(\alpha_{t-1}|Y_{t-1})$. For an appropriate Gaussian importance density $g(\alpha_t|\alpha_{t-1}, Y_t)$ we require estimates for $\mathrm{E}(\alpha_t|Y_t)$ and $\mathrm{Var}(\alpha_t|Y_t)$. We can obtain these via the extended Kalman filter or the unscented filter in combination with estimates such as $\bar{a}^+_{t-1|t-1}$ and $\bar{P}^+_{t-1|t-1}$. When considering the extended Kalman filter (10.4), we require estimates for $K_t = M_t F_t^{-1}$ and F_t in (10.4). For the unscented filter we require similar quantities but these are denoted by $P_{\alpha v,t}$, instead of M_t, and $P_{vv,t}$, instead of F_t; see equations (10.10) and (10.11). For the set of particles $\alpha^{(i)}_{t-1}$, $i = 1, \ldots, N$, estimators for $M_t = \mathrm{Cov}(\alpha_t, y_t|Y_{t-1})$ and $F_t = \mathrm{Var}(y_t|Y_{t-1})$ are obviously given by

$$\bar{M}^+_t = N^{-1} \sum_{i=1}^N \left(a^{+(i)}_t - \bar{a}^+_t\right)\left(v^{+(i)}_t - \bar{v}^+_t\right)',$$

$$\bar{F}^+_t = N^{-1} \sum_{i=1}^N \left(v^{+(i)}_t - \bar{v}^+_t\right)\left(v^{+(i)}_t - \bar{v}^+_t\right)',$$

where

$$a^{+(i)}_t = T_{t-1}(\alpha^{(i)}_{t-1}) + R_{t-1}(\alpha^{(i)}_{t-1})\eta^{(i)}_{t-1}, \quad \eta^{(i)}_{t-1} \sim \left[0, Q_{t-1}(\alpha^{(i)}_{t-1})\right],$$

$$\bar{a}^+_t = N^{-1} \sum_{i=1}^N a^{+(i)}_t,$$

and

$$v^{+(i)}_t = y_t - \mathrm{E}[y_t|a^{*(i)}_t, \varepsilon^{(i)}_t], \quad \varepsilon^{(i)}_t \sim \left[0, H_t(a^{*(i)}_t)\right], \quad \bar{v}^+_t = N^{-1} \sum_{i=1}^N v^{+(i)}_t,$$

with $a^{*(i)}_t = T_{t-1}(\alpha^{(i)}_{t-1})$ and expectation $\mathrm{E}(y_t|\alpha_t, \varepsilon_t)$ is with respect to $p[y_t|Z_t(\alpha_t), \varepsilon_t]$ for $\alpha_t = a^{*(i)}_t$ and $\varepsilon_t = \varepsilon^{(i)}_t$, for $i = 1, \ldots, N$. The variance matrix $H_t(\alpha_t)$ is with respect to $p(\varepsilon_t)$ or is a first-order Taylor approximation to the variance of $p(\varepsilon_t)$ at $\alpha_t = a^{*(i)}_t$. For the extended Kalman filter update in the particle filter, we define

$$\bar{K}^+_t = \bar{M}^+_t \bar{F}^{+\,-1}_t,$$

and compute

$$\bar{a}^+_{t|t} = N^{-1} \sum_{i=1}^N a^{+(i)}_{t|t}, \qquad \bar{P}^+_{t|t} = N^{-1} \sum_{i=1}^N \left(a^{+(i)}_{t|t} - \bar{a}^+_{t|t}\right)\left(a^{+(i)}_{t|t} - \bar{a}^+_{t|t}\right)',$$

where

$$a_{t|t}^{+(i)} = a_t^{+(i)} + \bar{K}_t^{+} v_t^{+(i)}, \qquad i = 1, \ldots, N.$$

We can set the importance density equal to

$$g(\alpha_t|\alpha_{t-1}, Y_t) = \mathrm{N}\left(a_{t|t}^{+(i)}, \bar{P}_{t|t}^{+}\right), \tag{12.18}$$

which we expect to be an accurate approximation of $p(\alpha_t|\alpha_{t-1}, Y_t)$. Next, the particle filter steps of the algorithm in Subsection 12.3.4 can be carried out with $g(\alpha_t|\alpha_{1:t-1}, Y_t) = g(\alpha_t|\alpha_{t-1}, Y_t)$ in step (i) given by (12.18). Alternative estimates for $\bar{a}_{t|t}^{+}$ and $\bar{P}_{t|t}^{+}$ can be obtained by considering weighted sample averages based on normalised weights $p(y_t|\alpha_t)$ evaluated at $\alpha_t = a_{t|t}^{+(i)}$ for $i = 1, \ldots, N$.

The mean, variance and covariance estimates as given above are subject to Monte Carlo error and require computational effort. An alternative is to use the extended or unscented Kalman filter approximations to obtain these estimates. In the case of the extended filter, the mean and variance estimates $a_{t|t}^{+}$ and $P_{t|t}^{+}$ can be obtained by $a_{t|t}$ and $P_{t|t}$, respectively, from (10.4) with

$$a_t = N^{-1} \sum_{i=1}^{N} T_{t-1}(\alpha_{t-1}^{(i)}),$$

and

$$P_t = N^{-1} \sum_{i=1}^{N} \left[T_{t-1}(\alpha_{t-1}^{(i)}) - a_t\right] \left[T_{t-1}(\alpha_{t-1}^{(i)}) - a_t\right]' \\ + R_{t-1}(\alpha_{t-1}^{(i)}) Q_{t-1}(\alpha_{t-1}^{(i)}) T_{t-1}(\alpha_{t-1}^{(i)})'.$$

In the case of the unscented filter, the mean and variance estimates $a_{t|t}^{+}$ and $P_{t|t}^{+}$ can be obtained by $a_{t|t}$ from (10.10) and $P_{t|t}$ from (10.11), respectively, where a_t and P_t can be computed as above.

The importance density (12.18) can be incorporated in both algorithms of Subsections 12.3.4 and 12.5.1. Step (i) of the algorithm in Subsection 12.3.4 can directly be based on (12.18). In the algorithm in Subsection 12.5.1, $\mu_t^{(i)}$ can be redefined in step (i) as a draw from importance density (12.18). Such modified algorithms are expected to be more effective since the information of y_t is taken into account when a new state variable α_t is generated.

12.6.2 The local regression filter

This filter can be used for the class of non-Gaussian state space models of the form

$$\begin{aligned} y_t &= Z_t \alpha_t + \varepsilon_t, & \varepsilon_t &\sim p(\varepsilon_t), \\ \alpha_{t+1} &= T_t \alpha_t + R_t \eta_t, & \eta_t &\sim p(\eta_t), \end{aligned} \tag{12.19}$$

where $\mathrm{E}(\varepsilon_t) = \mathrm{E}(\eta_t) = 0$, $\mathrm{V}(\varepsilon_t) = H_t$ and $\mathrm{V}(\eta_t) = Q_t$ for $t = 1, 2, \ldots$; matrices Z_t, T_t and R_t are assumed known. As before we would ideally like to take the importance density $g(\alpha_t|\alpha_{1:t-1}, Y_t)$ equal to $p(\alpha_t|\alpha_{t-1}, y_t)$ but this is not normally available in an analytical form. A further complication is that for many practical cases the dimension of α_t is greater than that of η_{t-1} so the conditional density of α_t given α_{t-1} is singular. For this reason it is more convenient to work with $(\alpha_{1:t-1}, \eta_{t-1})$ than working with $\alpha_{1:t}$ throughout. The idea behind the method is to obtain an importance density by approximating the conditional density of η_{t-1} given $\alpha_{1:t-1}$ and Y_t by $\mathrm{N}(b_t, c_t)$ where $b_t = \mathrm{E}(\eta_{t-1}|\alpha_{t-1}, y_t)$ and $c_t = \mathrm{V}(\eta_{t-1}|\alpha_{t-1}, y_t)$. The approximation does not need to be highly accurate. The advantage of the approach is that it takes the values of y_t explicitly into account in the selection of α_t.

Transform from $\alpha_{1:t}$ to $(\alpha_{1:t-1}, \eta_{t-1})$ and define $x_t^*(\alpha_{1:t-1}, \eta_{t-1}) = x_t(\alpha_{1:t})$. Analogously to (12.5) and (12.6), we have

$$\begin{aligned}\bar{x}_t &= \mathrm{E}(x_t^*(\alpha_{1:t-1}, \eta_{t-1})|Y_t) \\ &= \int x_t^*(\alpha_{1:t-1}, \eta_{t-1}) p(\alpha_{1:t-1}, \eta_{t-1}|Y_t) \mathrm{d}(\alpha_{1:t-1}, \eta_{t-1}) \\ &= \frac{1}{p(Y_t)} \mathrm{E}_g[x_t^*(\alpha_{1:t-1}, \eta_{t-1})\tilde{w}_t], \end{aligned} \qquad (12.20)$$

where E_g denotes expectation with respect to importance density $g(\alpha_{1:t-1}, \eta_{t-1}|Y_t)$ and

$$\tilde{w}_t = \frac{p(\alpha_{1:t-1}, \eta_{t-1}, Y_t)}{g(\alpha_{1:t-1}, \eta_{t-1}|Y_t)}. \qquad (12.21)$$

Now

$$\begin{aligned} p(\alpha_{1:t-1}, \eta_{t-1}, Y_t) &= p(\alpha_{1:t-1}, Y_{t-1}) p(\eta_{t-1}, y_t|\alpha_{1:t-1}, Y_{t-1}) \\ &= p(\alpha_{1:t-1}, Y_{t-1}) p(\eta_{t-1}) p(y_t|\alpha_t), \end{aligned}$$

with $\alpha_t = T_{t-1}\alpha_{t-1} + R_{t-1}\eta_{t-1}$, due to the Markovian structure of model (12.1) and (12.2). Analogously to (12.12), for particle filtering we must have

$$g(\alpha_{1:t-1}, \eta_{t-1}|Y_t) = g(\alpha_{1:t-1}|Y_{t-1}) g(\eta_{t-1}|\alpha_{1:t-1}, Y_t).$$

Assuming that importance densities $g(\cdot)$ have similar Markovian structure to that of model (12.1) and (12.2) we take $g(\eta_{t-1}|\alpha_{1:t-1}, Y_t) = g(\eta_{t-1}|\alpha_{t-1}, y_t)$. We therefore have from (12.21)

$$\tilde{w}_t = \frac{p(\alpha_{1:t-1}, Y_{t-1})}{g(\alpha_{1:t-1}|Y_{t-1})} \frac{p(\eta_{t-1}) p(y_t|\alpha_t)}{g(\eta_{t-1}|\alpha_{t-1}, y_t)}.$$

The Jacobians of the transformation from $\alpha_{1:t-1}, Y_{t-1}$ to $\alpha_{1:t-2}, \eta_{t-2}, Y_{t-1}$ cancel out so

$$\tilde{w}_t = \tilde{w}_{t-1} \frac{p(\eta_{t-1}) p(y_t|\alpha_t)}{g(\eta_{t-1}|\alpha_{t-1}, y_t)}, \qquad (12.22)$$

with $\alpha_t = T_{t-1}\alpha_{t-1} + R_{t-1}\eta_{t-1}$.

Putting $x^*(\alpha_{1:t-1}, \eta_{t-1}) = 1$ in (12.20) gives $p(Y_t) = \mathrm{E}_g(\tilde{w}_t)$ so we have

$$\bar{x}_t = \frac{\mathrm{E}_g[x^*(\alpha_{1:t-1}, \eta_{t-1})\tilde{w}_t]}{\mathrm{E}_g[\tilde{w}_t]}. \tag{12.23}$$

Taking a random sample $\eta_{t-1}^{(i)}$ from $g(\eta_{t-1})$ and computing $\alpha_t^{(i)} = T_{t-1}\alpha_{t-1}^{(i)} + R_{t-1}\eta_{t-1}^{(i)}$ for $i = 1, \ldots, N$, our estimate of x_t is

$$\hat{x}_t = \sum_{i=1}^{N} x_t(\alpha_{1:t}^{(i)}) w_t^{(i)},$$

as in (12.9), where

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^{N} \tilde{w}_t^{(i)}}, \qquad i = 1, \ldots, N, \tag{12.24}$$

with the recursion for $\tilde{w}_t^{(i)}$ given by

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-1}^{(i)} \frac{p(\eta_{t-1}^{(i)}) p(y_t|\alpha_t^{(i)})}{g(\eta_{t-1}^{(i)}|\alpha_{t-1}^{(i)}, y_t)}.$$

To construct the importance density $g(\eta_{t-1}|\alpha_{t-1}, y_t)$ we postulate that for $\alpha_{1:t-1}$ and Y_{t-1} fixed, η_{t-1} and y_t are generated by the linear Gaussian model

$$\begin{aligned} \eta_{t-1} &\sim \mathrm{N}(0, Q_{t-1}), & \alpha_t &= T_{t-1}\alpha_{t-1} + R_{t-1}\eta_{t-1}, \\ \varepsilon_t &\sim \mathrm{N}(0, H_t), & y_t &= Z_t\alpha_t + \varepsilon_t. \end{aligned}$$

We then choose η_{t-1} from the conditional distribution of η_{t-1} given α_{t-1} and y_t; this is a normal distribution which we denote by $N(b_t, c_t)$.

We have

$$\mathrm{E}(y_t|\alpha_{t-1}) = Z_t T_{t-1}\alpha_{t-1}, \qquad \mathrm{V}(y_t|\alpha_{t-1}) = W_t = Z_t R_{t-1} Q_{t-1} R_{t-1}' Z_t' + H_t.$$

Further, $\mathrm{E}(\eta_{t-1}y_t') = Q_{t-1}R_{t-1}'Z_t'$. It follows from elementary regression theory that

$$\begin{aligned} b_t &= \mathrm{E}(\eta_{t-1}|\alpha_{t-1}, y_t) = Q_{t-1}R_{t-1}'Z_t'W_t^{-1}(y_t - Z_t T_{t-1}\alpha_{t-1}), \\ c_t &= \mathrm{V}(\eta_{t-1}|\alpha_{t-1}, y_t) = Q_{t-1} - Q_{t-1}R_{t-1}'Z_t'W_t^{-1}Z_t R_{t-1}Q_{t-1}. \end{aligned} \tag{12.25}$$

The simulation steps are given by the following algorithm:

(i) Draw independent values $\eta_{t-1}^{(1)}, \ldots, \eta_{t-1}^{(N)}$ from $\mathrm{N}(b_t, c_t)$ and compute $\alpha_t^{(i)} = T_{t-1}\alpha_{t-1}^{(i)} + R_{t-1}\eta_{t-1}^{(i)}$ for $i = 1, \ldots, N$.

(ii) Compute normalised weights $w_t^{(1)}, \ldots, w_t^{(N)}$ as in (12.24) and resample with replacement using these weights.

(iii) Relabelling the resampled values as $\alpha_t^{(i)}$ and resetting the weights as $w_t^{(i)} = 1/N$, compute $\bar{x}_t = \frac{1}{N}\sum_{i=1}^{N} x(\alpha_{1:t})$.

12.6.3 The mode equalisation filter

This filter is intended for models of the form (9.1). We shall aim at obtaining a more accurate Gaussian approximation to $p(\eta_{t-1}|\alpha_{t-1}, y_t)$ than was used for the local regression filter. We do this by employing the method used in Section 10.6 for smoothing, that is, we choose the Gaussian approximating density which has the same mode as $p(\eta_{t-1}|\alpha_{t-1}, y_t)$. The construction of a similar importance density based on the mode is also explored by Doucet, Godsill and Andrieu (2000) and Cappé, Moulines and Rydén (2005, Chapter 7).

Instead of an importance density of the form $g(\eta_{t-1}|\alpha_{t-1}, y_t)$ we introduce a modified version $\tilde{y}_t$ of y_t and take as our importance density $g(\eta_{t-1}|\alpha_{t-1}, \tilde{y}_t)$ where η_{t-1} and $\tilde{y}_t$ are generated by the linear Gaussian model

$$\begin{aligned} \eta_{t-1} &\sim N(0, \tilde{Q}_{t-1}), & \varepsilon_t &\sim N(0, \tilde{H}_t), \\ \alpha_t &= T_{t-1}\alpha_{t-1} + R_{t-1}\eta_{t-1}, & \tilde{y}_t &= Z_t\alpha_t + \varepsilon_t. \end{aligned}$$

We shall determine $\tilde{y}_t$, $\tilde{Q}_{t-1}$ and $\tilde{H}_t$ so that $g(\eta_{t-1}|\alpha_{t-1}, \tilde{y}_t)$ and $p(\eta_{t-1}|\alpha_{t-1}, y_t)$ have the same mode to a good enough approximation.

The mode of density $g(\eta_{t-1}|\alpha_{t-1}, \tilde{y}_t)$ is the solution of the equation

$$\frac{\partial \log g(\eta_{t-1}|\alpha_{t-1}, \tilde{y}_t)}{\partial \eta_{t-1}} = 0.$$

Since $\log g(\eta_{t-1}|\alpha_{t-1}, \tilde{y}_t) = \log g(\eta_{t-1}, \tilde{y}_t|\alpha_{t-1}) - \log g(\tilde{y}_t|\alpha_{t-1})$, it can also be obtained as

$$\frac{\partial \log g(\eta_{t-1}, \tilde{y}_t|\alpha_{t-1})}{\partial \eta_{t-1}} = 0.$$

We have

$$\log g(\eta_{t-1}, \tilde{y}_t|\alpha_{t-1}) = \text{constant} - \frac{1}{2}[\eta_{t-1}'\tilde{Q}_t^{-1}\eta_{t-1} + (\tilde{y}_t - Z_t\alpha_t)'\tilde{H}_t^{-1}(\tilde{y}_t - Z_t\alpha_t)],$$

where $\alpha_t = T_{t-1}\alpha_{t-1} + R_{t-1}\eta_{t-1}$. Differentiating and equating to zero gives

$$-\tilde{Q}_{t-1}^{-1}\eta_{t-1} + R_t'Z_t'\tilde{H}_t^{-1}(\tilde{y}_t - Z_t\alpha_t) = 0. \tag{12.26}$$

The mode of a Gaussian density is equal to the mean. Therefore, at first sight the solution for η_{t-1} of (12.26) appears inconsistent with the expression for b_t in (12.25). For the two solutions to be equal we need to have

$$\begin{aligned}&\tilde{Q}_{t-1}R'_{t-1}Z'_t(Z_tR_{t-1}\tilde{Q}_{t-1}R'_{t-1}Z'_t+\tilde{H}_t)^{-1}\\&\quad=(\tilde{Q}_{t-1}^{-1}+R'_{t-1}Z'_t\tilde{H}_t^{-1}Z_tR_{t-1})^{-1}R'_{t-1}Z'_t\tilde{H}_t^{-1}.\end{aligned}$$

The equality can be verified by premultiplying the equation by $\tilde{Q}_{t-1}^{-1}+R'_{t-1}Z'_t\tilde{H}_t^{-1}Z_tR_{t-1}$ and postmultiplying by $Z_tR_{t-1}\tilde{Q}_{t-1}R'_{t-1}Z'_t+\tilde{H}_t$.

Denote $Z_t\alpha_t$ by θ_t. For the model

$$p(y_t|\theta_t),\qquad \theta_t=Z_t\alpha_t,\qquad \alpha_t=T_{t-1}\alpha_{t-1}+R_{t-1}\eta_{t-1},\qquad \eta_{t-1}\sim p(\eta_{t-1}),$$

the mode of η_{t-1} given α_{t-1} and y_t is the solution of the equation

$$\frac{\partial\log p(\eta_{t-1},y_t|\alpha_{t-1})}{\partial\eta_{t-1}}+\frac{\partial\log p(\eta_{t-1})}{\partial\eta_{t-1}}+R'_{t-1}Z'_t\frac{\partial\log p(y_t|\theta_t)}{\partial\theta_t}=0.\qquad(12.27)$$

We want to find $\tilde{y}_t$, $\tilde{Q}_{t-1}$ and $\tilde{H}_t$ so that the solutions of (12.26) and (12.27) are the same; we do this by iteration based on linearisation. Suppose that $\tilde{\eta}_{t-1}$ is a trial value of η_{t-1}. Let

$$\tilde{\theta}_t=Z_t(T_{t-1}\alpha_{t-1}+R_{t-1}\tilde{\eta}_{t-1}),\qquad \dot{h}_t=-\left.\frac{\partial\log p(y_t|\theta_t)}{\partial\theta_t}\right|_{\theta_t=\tilde{\theta}_t},$$

$$\ddot{h}_t=-\left.\frac{\partial^2\log p(y_t|\theta_t)}{\partial\theta_t\partial\theta'_t}\right|_{\theta_t=\tilde{\theta}_t},$$

and take

$$\tilde{H}_t=\ddot{h}_t^{-1},\qquad \tilde{y}_t=\tilde{\theta}_t-\ddot{h}_t^{-1}\dot{h}_t.\qquad(12.28)$$

The linearised form of

$$\frac{\partial\log p(y_t|\theta_t)}{\partial\theta_t},$$

at $\theta_t=\tilde{\theta}_t$ is $-\dot{h}_t-\ddot{h}_t(\theta_t-\tilde{\theta}_t)=\tilde{H}_t^{-1}(\tilde{y}_t-\theta_t)$. Substituting this in (12.27) and comparing the result with (12.26), we see that we have achieved the linearised form required for the observation term of (12.27).

To linearise the state term of (12.27) when $p(\eta_{t-1})$ is non-Gaussian, we proceed as in Section 11.6. We assume that $\eta_{t-1,i}$, the ith element of η_{t-1}, is independent of the other elements in η_{t-1} and that its density is a function of $\eta_{t-1,i}^2$, for $i=1,\dots,r$. Although these assumptions appear restrictive, they enable us to deal with the important case in which the error densities have heavy tails. Let $r_{t-1,i}(\eta_{t-1,i}^2)=-2\log p(\eta_{t-1,i})$, and let

$$\dot{r}_{t-1,i}=\left.\frac{\partial r_{t-1,i}(\eta_{t-1,i}^2)}{\partial\eta_{t-1,i}^2}\right|_{\eta_{t-1}=\tilde{\eta}_{t-1}},$$

for a trial value $\tilde{\eta}_{t-1}$ of η_{t-1}. Then the linearised form of the state term (12.26) is

$$-\sum_{i=1}^{r} \dot{r}_{t-1,i}\eta_{t-1,i}. \qquad (12.29)$$

Putting $\tilde{Q}_{t-1}^{-1} = \mathrm{diag}(\dot{r}_{t-1,1}, \dots, \dot{r}_{t-1,r})$ we see that (12.29) has the same form as the state component of (12.26).

Starting with the trial value $\tilde{\eta}_{t-1}$ of η_{t-1} we use the solution

$$\eta_{t-1} = (\tilde{Q}_{t-1}^{-1} + R_{t-1}'Z_t'\tilde{H}_t^{-1}Z_tR_{t-1})^{-1}R_{t-1}'Z_t'\tilde{H}_t^{-1}(\tilde{y}_t - Z_tT_{t-1}\alpha_{t-1}),$$

of (12.26) to obtain a new value of η_{t-1} which we use as the next trial value. We repeat the process and continue until reasonable convergence is achieved.

12.7 Rao–Blackwellisation

12.7.1 Introduction

Suppose that we can partition the state vector into two components

$$\alpha_t = \begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \end{pmatrix},$$

where α_{1t} and α_{2t} are such that the integration with respect to α_{1t} given α_{2t} can be performed analytically. For example, consider the univariate local level model (2.3) with a state equation variance η_t which varies over time. This model is given by

$$y_t = \mu_t + \varepsilon_t, \qquad \mu_{t+1} = \mu_t + \eta_t, \qquad (12.30)$$

where $\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ and $\eta_t \sim \mathrm{N}(0, \exp h_t)$ with

$$h_{t+1} = (1-\phi)h^* + \phi h_t + \zeta_t, \qquad \zeta_t \sim \mathrm{N}(0, \sigma_\zeta^2), \qquad (12.31)$$

and known fixed parameters σ_ε^2, h^*, ϕ and σ_ζ^2. The model is a variant of the local level model (2.3) in which η_t is modelled similarly to y_t of the stochastic volatility model (9.26) and (9.27) with $\mu = 0$ and $\sigma^2 \exp\theta_t = \exp h_t$. Since both μ_t and h_t evolve stochastically, we can place them both in the state vector α_t and take $\alpha_{1t} = \mu_t$ and $\alpha_{2t} = h_t$. It follows that for a given α_{2t}, this local level model reduces to a standard linear Gaussian model with known time-varying variances. The standard Kalman filter and smoothing methods of Chapter 4 can be employed for the estimation of α_{1t} given α_{2t}, or, in the wording of this section, for the analytical integration of α_{1t}.

We proceed as follows. Choose a sample $h_1^{(i)}, \dots, h_n^{(i)}$, for $i = 1, \dots, N$, using one of the techniques described in Section 12.4 applied to (12.31). Next, apply the Kalman filter and smoother to (12.30) with $\mathrm{Var}(\eta_t) = \exp h_t^{(i)}$ for $i = 1, \dots, N$.

By using this approach, we only need to simulate h_t instead of having to simulate both μ_t and h_t as is required for the general analysis of nonlinear models described in Section 12.3.

This device for reducing the amount of simulation required by analytical integration of one component of a state vector for given values of another component is called *Rao-Blackwellisation.*

12.7.2 The Rao–Blackwellisation technique

The gain in efficiency from Rao–Blackwellisation arises from the following basic result in estimation theory. Suppose that a vector z is a function of the random vectors x and y with mean $\mathrm{E}(z) = \mu$. Let $z^* = \mathrm{E}(z|x)$; then $\mathrm{E}(z^*) = \mu$ and

$$\begin{aligned}\mathrm{Var}(z) &= \mathrm{E}\left[(z-\mu)(z-\mu)'\right] \\ &= \mathrm{E}\left\{\mathrm{E}\left[(z-z^*+z^*-\mu)(z-z^*+z^*-\mu)'\right]|x\right\} \\ &= E\left[\mathrm{Var}(z|x)\right] + \mathrm{Var}(z^*).\end{aligned}$$

Since $\mathrm{Var}(z|x)$ is non-negative definite, it follows that $\mathrm{Var}(z^*)$ is equal to or smaller in the matrix sense than $\mathrm{Var}(z)$. If z is regarded as an estimate of μ and $\mathrm{Var}(z^*)$ is strictly smaller than $\mathrm{Var}(z)$, then its conditional expectation z^* is an improved estimate.

Strictly speaking, the Rao–Blackwell theorem is concerned with the special case where x is a sufficient statistic for an unknown parameter; see, for example, Rao (1973, §5a.2(iii)) and Lehmann (1983, Theorem 6.4). Since in applications using the state space model, variable x is not a sufficient statistic, the use of the term 'Rao–Blackwellisation' is to some extent inappropriate; nevertheless since its use in the literature is well established we shall continue to use it here.

Denote the sets $\{\alpha_{1,1}, \ldots, \alpha_{1,t}\}$ and $\{\alpha_{2,1}, \ldots, \alpha_{2,t}\}$ by $\alpha_{1,1:t}$ and $\alpha_{2,1:t}$, respectively. As in Section 12.3, let x_t be an arbitrary function of α_t and suppose that we wish to estimate the conditional mean $\bar{x}_t$ given by

$$\begin{aligned}\bar{x}_t \;=\; & \mathrm{E}\left[x_t(\alpha_{1:t}|Y_t)\right] = \mathrm{E}\left[x_t(\alpha_{1,1:t}, \alpha_{2,1:t})|Y_t)\right] \\ & = \int x_t(\alpha_{1,1:t}, \alpha_{2,1:t}) p_t(\alpha_{1,1:t}, \alpha_{2,1:t}|Y_t)\, \mathrm{d}\alpha_{1,1:t}\, \mathrm{d}\alpha_{2,1:t}.\end{aligned}$$

Let

$$h_t(\alpha_{2,1:t}) = \int x_t(\alpha_{1,1:t}, \alpha_{2,1:t}) p_t(\alpha_{1,1:t}|\alpha_{2,1:t}, Y_t)\, \mathrm{d}\alpha_{1,1:t}. \tag{12.32}$$

We then have

$$\bar{x}_t = \int h_t(\alpha_{2,1:t}) p_t(\alpha_{2,1:t}|Y_t)\, \mathrm{d}\alpha_{2,1:t}. \tag{12.33}$$

We assume that $h_t(\alpha_{2,1:t})$ can be calculated analytically for given values of $\alpha_{2,1:t}$. Then (12.33) has the same form as (12.5) so similar methods can be used for evaluation of (12.33) as were used in Section 12.2.

Let $g(\alpha_{2,1:t}|Y_t)$ be an importance density chosen to be close to $p(\alpha_{2,1:t}|Y_t)$ and choose a random sample $\alpha_{2,1:t}^{(1)}, \dots, \alpha_{2,1:t}^{(N)}$ from density $g(\alpha_{2,1:t}|Y_t)$. Then as in Section 12.2, $\bar{x}_t$ can be estimated by

$$\hat{x}_t = \sum_{i=1}^{N} h_t(\alpha_{2,1:t}^{(i)}) w_t^{(i)}, \tag{12.34}$$

with

$$\tilde{w}_t^{(i)} = \frac{p_t(\alpha_{2,1:t}^{(i)}, Y_t)}{q_t(\alpha_{2,1:t}^{(i)}|Y_t)}, \qquad \text{and} \qquad w_t^{(i)} = \tilde{w}_t^{(i)} \Big/ \sum_{j=1}^{N} \tilde{w}_t^{(j)}.$$

Here, appropriate particle filtering methods can be employed as in Section 12.2.

Since $h_t(\alpha_{2,1:t})$ is calculated analytically for each sample value $\alpha_{2,1:t} = \alpha_{2,1:t}^{(i)}$ for $i = 1, \dots, N$, for example by means of the Kalman filter, computational gains are typically high. Although Rao–Blackwellisation can be used generally, it is specifically advantageous for high-dimensional models, where standard particle filtering can be computationally costly. This point is discussed by Doucet, De Freitas and Gordon (2001, pp. 499–515) who consider an example where the state vector has the huge dimension of 2^{100}. Further details are discussed in Doucet, Godsill and Andrieu (2000) and Maskell (2004).

13 Bayesian estimation of parameters

13.1 Introduction

The parameters of a state space model can be divided into two classes, the state parameters and the additional parameters. For example, in the local level model (2.3) the state parameters are functions of $\alpha_1, \ldots, \alpha_n$ and the additional parameters are σ_ε^2 and σ_η^2. There are two methods of estimation of these parameters, classical analysis and Bayesian analysis; these arise from two different theories of probability. Classical analysis comes from a theory in which parameters are fixed and observations are random variables; Bayesian analysis comes from a theory in which state parameters are random variables and the observations are fixed. In state space analysis, state estimates are the same whether classical analysis or Bayesian analysis is employed by Lemmas 1 to 4 of Chapter 4. However, different treatments are needed for the estimation of additional parameters. classical treatments are based on fixed parameters and we have shown in previous chapters how to use the method of maximum likelihood for their estimation. For Bayesian treatments it turns out that we need to use simulation-based methods for the estimation of additional parameters, even for linear models. Thus it seemed logical to delay the Bayesian treatment of linear models until after simulation methods for nonlinear and non-Gaussian models had been dealt with. It was then natural to run on with the Bayesian treatment of nonlinear and non-Gaussian models. This explains why we did not attempt to deal with the Bayesian estimation of additional parameters in Part I of the book. This chapter deals with the Bayesian analysis of additional parameters as well as state parameters.

A Bayesian treatment of state space models is provided by the monograph of West and Harrison (1989, 1997). Other previous work on Bayesian analysis of non-Gaussian state space models has been based mainly on Markov chain Monte Carlo (MCMC) methods; we note in particular here the contributions by Carlin, Polson and Stoffer (1992), Frühwirth-Schnatter (1994, 2004), Shephard (1994b), Carter and Kohn (1994, 1996, 1997), Shephard and Pitt (1997), Cargnoni, Muller and West (1997) and Gamerman (1998). General accounts of Bayesian methodology and computation are given by Bernardo and Smith (1994), Gelman, Carlin, Stern and Rubin (1995) and Frühwirth-Schnatter (2006).

We first develop the analysis of the linear Gaussian state space model by constructing importance samples of additional parameters. We then show how to combine these with Kalman filter and smoother outputs to obtain the estimates of state parameters required. We first do this for a proper prior in Subsection 13.2.1 and then extend to non-informative priors in Subsection 13.2.2. The discussion is extended to nonlinear and non-Gaussian models in Subsection 13.3. In Section 13.4 a brief description is given of the alternative simulation technique, Markov chain Monte Carlo (MCMC) methods. Although we prefer to use importance sampling methods for the problems considered in this book, MCMC methods are also used in time series applications.

13.2 Posterior analysis for linear Gaussian model

13.2.1 Posterior analysis based on importance sampling

Let ψ denote the vector of additional parameters. We consider a Bayesian analysis for the situation where the parameter vector ψ is not fixed and known; instead, we treat ψ as a random vector with a known prior density $p(\psi)$, which to begin with we take as a proper prior, leaving the non-informative case until later. For discussions of choice of prior in a general Bayesian analysis see Gelman, Carlin, Stern and Rubin (1995) and Bernardo and Smith (1994). The problems we shall consider amount essentially to the estimation of the posterior mean of a function $x(\alpha)$ of the stacked state vector α,

$$\bar{x} = \mathrm{E}[x(\alpha)|Y_n]. \tag{13.1}$$

Let $\bar{x}(\psi) = \mathrm{E}[x(\alpha)|\psi, Y_n]$ be the conditional expectation of $x(\alpha)$ given ψ and Y_n. We shall restrict consideration in this section to those functions $x(\alpha)$ for which $\bar{x}(\psi)$ can be readily calculated by the Kalman filter and smoother. This restricted class of functions still, however, includes many important cases such as the posterior mean and variance matrix of α_t and forecasts $\mathrm{E}(y_{n+j}|Y_n)$ for $j = 1, 2, \ldots$. The treatment of initialisation in Chapter 5 permits elements of the initial state vector α_1 to have either proper or diffuse prior densities.

We begin by attempting an analysis which is free of importance sampling. We have

$$\bar{x} = \int \bar{x}(\psi) p(\psi|Y_n)\, d\psi. \tag{13.2}$$

By Bayes theorem $p(\psi|Y_n) = Kp(\psi)p(Y_n|\psi)$ where K is the normalising constant defined by

$$K^{-1} = \int p(\psi) p(Y_n|\psi)\, d\psi. \tag{13.3}$$

We therefore have

$$\bar{x} = \frac{\int \bar{x}(\psi) p(\psi) p(Y_n|\psi)\, d\psi}{\int p(\psi) p(Y_n|\psi)\, d\psi}. \tag{13.4}$$

Now $p(Y_n|\psi)$ is the likelihood, which for the linear Gaussian model is calculated by the Kalman filter as shown in Section 7.2. In principle, simulation could be applied directly to formula (13.4) by drawing a random sample $\psi^{(1)}, \ldots, \psi^{(N)}$

from the distribution with density $p(\psi)$ and then estimating the numerator and denominatior of (13.4) by the sample means of $\bar{x}(\psi)p(Y_n|\psi)$ and $p(Y_n|\psi)$ respectively. However, this estimator is inefficient in cases of practical interest.

We provide a treatment based on the simulation method of importance sampling that we have explored in Chapter 11. Suppose that simulation from density $p(\psi|Y_n)$ is impractical. Let $g(\psi|Y_n)$ be an importance density which is as close as possible to $p(\psi|Y_n)$ while at the same time permitting simulation. From (13.2) we have

$$\begin{aligned}\bar{x} &= \int \bar{x}(\psi)\frac{p(\psi|Y_n)}{g(\psi|Y_n)}g(\psi|Y_n)\,d\psi \\ &= \mathrm{E}_g\left[\bar{x}(\psi)\frac{p(\psi|Y_n)}{g(\psi|Y_n)}\right] \\ &= K\mathrm{E}_g[\bar{x}(\psi)z^g(\psi,Y_n)] \end{aligned} \tag{13.5}$$

by Bayes theorem where E_g denotes expectation with respect to density $g(\psi|Y_n)$,

$$z^g(\psi,Y_n) = \frac{p(\psi)p(Y_n|\psi)}{g(\psi|Y_n)}, \tag{13.6}$$

and K is a normalising constant. By replacing $\bar{x}(\psi)$ by 1 in (13.5) we obtain

$$K^{-1} = \mathrm{E}_g[z^g(\psi,Y_n)],$$

so the posterior mean of $x(\alpha)$ can be expressed as

$$\bar{x} = \frac{\mathrm{E}_g[\bar{x}(\psi)z^g(\psi,Y_n)]}{\mathrm{E}_g[z^g(\psi,Y_n)]}. \tag{13.7}$$

This expression is evaluated by simulation. We choose random samples of N draws of ψ, denoted by $\psi^{(i)}$, from the importance density $g(\psi|Y_n)$ and estimate $\bar{x}$ by

$$\hat{x} = \frac{\sum_{i=1}^N \bar{x}\big(\psi^{(i)}\big)z_i}{\sum_{i=1}^N z_i}, \tag{13.8}$$

where

$$z_i = \frac{p\big(\psi^{(i)}\big)p\big(Y_n|\psi^{(i)}\big)}{g\big(\psi^{(i)}|Y_n\big)}. \tag{13.9}$$

As an importance density for $p(\psi|Y_n)$ we take its large sample normal approximation

$$g(\psi|Y_n) = \mathrm{N}(\hat{\mu},\hat{\Omega}),$$

where $\hat{\psi}$ is the solution to the equation

$$\frac{\partial \log p(\psi|Y_n)}{\partial \psi} = \frac{\partial \log p(\psi)}{\partial \psi} + \frac{\partial \log p(Y_n|\psi)}{\partial \psi} = 0, \tag{13.10}$$

and

$$\hat{\Omega}^{-1} = -\frac{\partial^2 \log p(\psi)}{\partial \psi \partial \psi'} - \left.\frac{\partial^2 \log p(Y_n|\psi)}{\partial \psi \partial \psi'}\right|_{\psi=\hat{\psi}}. \tag{13.11}$$

For a discussion of this large sample approximation to $p(\psi|Y_n)$ see Gelman, Carlin, Stern and Rubin (1995, Chapter 4) and Bernardo and Smith (1994, §5.3). Since $p(Y_n|\psi)$ can easily be computed by the Kalman filter for $\psi = \psi^{(i)}$, $p(\psi)$ is given and $g(\psi|Y_n)$ is Gaussian, the value of z_i is easy to compute. The draws for $\psi^{(i)}$ are independent and therefore $\hat{x}$ converges probabilistically to $\bar{x}$ as $N \to \infty$ under very general conditions.

The value $\hat{\psi}$ is computed iteratively by an obvious extension of the technique of maximum likelihood estimation as discussed in Chapter 7 while the second derivatives can be calculated numerically. Once $\hat{\psi}$ and $\hat{\Omega}$ are computed, it is straightforward to generate samples from $g(\psi|Y_n)$ by use of a standard normal random number generator. Where needed, efficiency can be improved by the use of antithetic variables, which we discuss in Subsection 11.4.3. For example, for each draw $\psi^{(i)}$ we could take another value $\tilde{\psi}^{(i)} = 2\hat{\psi} - \psi^{(i)}$, which is equiprobable with $\psi^{(i)}$. The use of $\psi^{(i)}$ and $\tilde{\psi}^{(i)}$ together introduces balance in the sample.

The posterior mean of the parameter vector ψ is $\bar{\psi} = \mathrm{E}(\psi|Y_n)$. An estimate $\tilde{\psi}$ of $\bar{\psi}$ is obtained by putting $\bar{x}(\psi^{(i)}) = \psi^{(i)}$ in (13.8) and taking $\tilde{\psi} = \hat{x}$. Similarly, an estimate $\tilde{\mathrm{V}}(\psi|Y_n)$ of the posterior variance matrix $\mathrm{Var}(\psi|Y_n)$ is obtained by putting $\bar{x}(\psi^{(i)}) = \psi^{(i)}\psi^{(i)\prime}$ in (13.8), taking $\tilde{S} = \hat{x}$ and then taking $\tilde{\mathrm{V}}(\psi|Y_n) = \tilde{S} - \tilde{\psi}\tilde{\psi}'$.

To estimate the posterior distribution function of an element ψ_1 of ψ, which is not necessarily the first element of ψ, we introduce the indicator function $I_1(\psi_1^{(i)})$ which equals one if $\psi_1^{(i)} \leq \psi_1$ and zero otherwise, where $\psi_1^{(i)}$ is the value of ψ_1 in the ith simulated value of ψ and ψ_1 is fixed. Then $\mathrm{F}(\psi_1|Y_n) = \Pr(\psi_1^{(i)} \leq \psi_1) = \mathrm{E}[I_1(\psi_1^{(i)})|Y_n]$ is the posterior distribution function of ψ_1. Putting $\bar{x}(\psi^{(i)}) = I_1(\psi^{(i)})$ in (13.8), we estimate $\mathrm{F}(\psi_1|Y_n)$ by $\tilde{\mathrm{F}}(\psi_1|Y_n) = \hat{x}$. This is equivalent to taking $\tilde{\mathrm{F}}(\psi_1|Y_n)$ as the sum of values of z_i for which $\psi_1^{(i)} \leq \psi_1$ divided by the sum of all values of z_i. Similarly, if δ is the interval $(\psi_1 - \frac{1}{2}d, \psi_1 + \frac{1}{2}d)$ where d is small and positive then we can estimate the posterior density of ψ_1 by $\tilde{p}(\psi_1|Y_n) = d^{-1}S^{\delta} / \sum_{i=1}^{N} z_i$ where S^{δ} is the sum of the values of z_i for which $\psi_1^{(i)} \in \delta$.

13.2.2 Non-informative priors

For cases where a proper prior is not available we may wish to use a noninformative prior in which we assume that the prior density is proportional to a specified function $p(\psi)$ in a domain of ψ of interest even though the integral $\int p(\psi)d\psi$ does not exist. For a discussion of noninformative priors see, for example, Chapters 3 and 4 of Gelman, Carlin, Stern and Rubin (1995). Where it exists, the posterior density is $p(\psi|Y_n) = Kp(\psi)p(Y_n|\psi)$ as in the proper prior case, so all the previous formulae apply without change. This is why we use the same symbol $p(\psi)$ for both cases even though in the noninformative case $p(\psi)$ is not a density. An important special case is the diffuse prior for which $p(\psi) = 1$ for all ψ.

13.3 Posterior analysis for a nonlinear non-Gaussian model

In this section we develop Bayesian techniques for estimating posterior means and posterior variance matrices of functions of the state vector for a nonlinear non-Gaussian model. We also show how to estimate posterior distribution and density functions of scalar functions of the state vector. It turns out that the basic ideas of importance sampling and antithetic variables developed for classical analysis in Chapter 11 can be applied with little essential change to the Bayesian case. Different considerations apply to questions regarding the posterior distribution of the parameter vector and we deal with these in Subsection 13.3.3. The treatment is based on the methods developed by Durbin and Koopman (2000).

13.3.1 Posterior analysis of functions of the state vector

We first obtain some basic formulae analogous to those derived in Section 11.2 for the classical case. Suppose that we wish to calculate the posterior mean $\bar{x} = \mathrm{E}[x(\alpha)|Y_n]$ of a function $x(\alpha)$ of the stacked state vector α given the stacked observation vector Y_n. As we shall show, this is a general formulation which enables us not only to estimate posterior means of quantities of interest such as the trend or seasonal, but also posterior variance matrices and posterior distribution functions and densities of scalar functions of the state. We shall estimate $\bar{x}$ by simulation techniques based on importance sampling and antithetic variables analogous to those developed in Chapter 11 for the classical case.

We have

$$\begin{aligned}\bar{x} &= \int x(\alpha)p(\psi, \alpha|Y_n)\, d\psi d\alpha \\ &= \int x(\alpha)p(\psi|Y_n)p(\alpha|\psi, Y_n)\, d\psi d\alpha. \qquad (13.12)\end{aligned}$$

As an importance density for $p(\psi|Y_n)$ we take its large sample normal approximation

$$g(\psi|Y_n) = \mathrm{N}(\hat{\psi}, \hat{V}),$$

where $\hat{\psi}$ is the solution of the equation

$$\frac{\partial \log p(\psi|Y_n)}{\partial \psi} = \frac{\partial \log p(\psi)}{\partial \psi} + \frac{\partial \log p(Y_n|\psi)}{\partial \psi} = 0, \qquad (13.13)$$

and

$$\hat{V}^{-1} = -\frac{\partial^2 \log p(\psi)}{\partial\psi\partial\psi'} - \frac{\partial^2 \log p(Y_n|\psi)}{\partial\psi\partial\psi'}\bigg|_{\psi=\hat{\psi}}. \tag{13.14}$$

For a discussion of this large sample approximation to $p(\psi|Y_n)$ see Gelman, Carlin, Stern and Rubin (1995, Chapter 4) and Bernardo and Smith (1994, §5.3).

Let $g(\alpha|\psi, Y_n)$ be a Gaussian importance density for α given ψ and Y_n which is obtained from an approximating linear Gaussian model in the way described in Chapter 11. From (13.12),

$$\begin{aligned}\bar{x} &= \int x(\alpha)\frac{p(\psi|Y_n)p(\alpha|\psi, Y_n)}{g(\psi|Y_n)g(\alpha|\psi, Y_n)}g(\psi|Y_n)g(\alpha|\psi, Y_n)\, d\psi d\alpha \\ &= \int x(\alpha)\frac{p(\psi|Y_n)g(Y_n|\psi)p(\alpha, Y_n|\psi)}{g(\psi|Y_n)p(Y_n|\psi)g(\alpha, Y_n|\psi)}g(\psi, \alpha|Y_n)\, d\psi d\alpha.\end{aligned}$$

By Bayes theorem,

$$p(\psi|Y_n) = Kp(\psi)p(Y_n|\psi),$$

in which K is a normalising constant, so we have

$$\begin{aligned}\bar{x} &= K\int x(\alpha)\frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)}\frac{p(\alpha, Y_n|\psi)}{g(\alpha, Y_n|\psi)}g(\psi, \alpha|Y_n)\, d\psi d\alpha \\ &= K\mathrm{E}_g\left[x(\alpha)z(\psi, \alpha, Y_n)\right], \end{aligned} \tag{13.15}$$

where E_g denotes expectation with respect to the importance joint density

$$g(\psi, \alpha|Y_n) = g(\psi|Y_n)g(\alpha|\psi, Y_n),$$

and where

$$z(\psi, \alpha, Y_n) = \frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)}\frac{p(\alpha, Y_n|\psi)}{g(\alpha, Y_n|\psi)}. \tag{13.16}$$

In this formula, $g(Y_n|\psi)$ is the likelihood for the approximating Gaussian model, which is easily calculated by the Kalman filter.

Taking $x(\alpha) = 1$ in (13.15) gives

$$K^{-1} = \mathrm{E}_g[z(\psi, \alpha, Y_n)],$$

so we have finally

$$\bar{x} = \frac{\mathrm{E}_g[x(\alpha)z(\psi, \alpha, Y_n)]}{\mathrm{E}_g[z(\psi, \alpha, Y_n)]}. \tag{13.17}$$

We note that (13.17) differs from the corresponding formula (11.8) in the classical inference case only in the replacement of $w(\alpha, Y_n)$ by $z(\psi, \alpha, Y_n)$ and the inclusion of ψ in the importance density $g(\psi, \alpha | Y_n)$.

In the important special case in which the state equation error η_t is $\mathrm{N}(0, Q_t)$, then α is Gaussian so we can write its density as $g(\alpha)$ and use this as the state density for the approximating model. This gives $p(\alpha, Y_n|\psi) = g(\alpha)p(Y_n|\theta, \psi)$ and $g(\alpha, Y_n|\psi) = g(\alpha)g(Y_n|\theta, \psi)$, where θ is the stacked vector of signals $\theta_t = Z_t\alpha_t$, so (13.16) simplifies to

$$z(\psi, \alpha, Y_n) = \frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)} \frac{p(Y_n|\theta, \psi)}{g(Y_n|\theta, \psi)}. \tag{13.18}$$

For cases where a proper prior is not available, we may wish to use a noninformative prior in which we assume that the prior density is proportional to a specified function $p(\psi)$ in a domain of ψ of interest even though the integral $\int p(\psi)d\psi$ does not exist. The posterior density, where it exists, is

$$p(\psi|Y_n) = Kp(\psi)p(Y_n|\psi),$$

which is the same as in the proper prior case, so all the previous formulae apply without change. This is why we can use the same symbol $p(\psi)$ in both cases even when $p(\psi)$ is not a proper density. An important special case is the diffuse prior for which $p(\psi) = 1$ for all ψ. For a general discussion of noninformative priors, see, for example, Gelman, Carlin, Stern and Rubin (1995, Chapters 2 and 3).

13.3.2 Computational aspects of Bayesian analysis

For practical computations based on these ideas we express the formulae in terms of variables that are as simple as possible as in Section 11.4, Subsections 11.5.3 and 11.5.5 for the classical analysis. This means that to the maximum feasible extent we employ formulae based on the disturbance terms $\eta_t = R_t'(\alpha_{t+1} - T_t\alpha_t)$ and $\varepsilon_t = y_t - \theta_t$ for $t = 1, \ldots, n$. By repeated substitution for α_t we first obtain $x(\alpha)$ as a function $x^*(\eta)$ of η. We then note that in place of (13.12) we obtain the posterior mean of $x^*(\eta)$,

$$\bar{x} = \int x^*(\eta)p(\psi|Y_n)p(\eta|\psi, Y_n)\, d\psi d\eta. \tag{13.19}$$

By reductions analogous to those above we obtain in place (13.17)

$$\bar{x} = \frac{\mathrm{E}_g[x^*(\eta)z^*(\psi, \eta, Y_n)]}{\mathrm{E}_g[z^*(\psi, \eta, Y_n)]}, \tag{13.20}$$

where

$$z^*(\psi, \eta, Y_n) = \frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)} \frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)}, \tag{13.21}$$

and E_g denotes expectation with respect to the importance density $g(\psi, \eta|Y_n)$.

Let $\psi^{(i)}$ be a random draw from the importance density for ψ, $g(\psi|Y_n) = \mathrm{N}(\hat{\psi}, \hat{V})$, where $\hat{\psi}$ satisfies (13.13) and $\hat{V}$ is given by (13.14), and let $\eta^{(i)}$ be a random draw from density $g(\eta|\psi^{(i)}, Y_n)$ for $i = 1, \ldots, N$. To obtain this we need an approximation to the mode $\hat{\eta}^{(i)}$ of density $g(\eta|\psi^{(i)}, Y_n)$ but this is rapidly obtained in a few iterations from the mode of $g(\eta|\hat{\psi}, Y_n)$. Let

$$x_i = x^*\big(\eta^{(i)}\big), \qquad z_i = z^*\big(\psi^{(i)}, \eta^{(i)}, Y_n\big), \tag{13.22}$$

and consider as an estimate of $\bar{x}$ the ratio

$$\hat{x} = \frac{\sum_{i=1}^N x_i z_i}{\sum_{i=1}^N z_i}. \tag{13.23}$$

The efficiency of this estimate can obviously be improved by the use of antithetic variables. For $\eta^{(i)}$ we can use the location and scale antithetics described in Subsection 11.4.3. Antithethics may not be needed for $\psi^{(i)}$ since $\hat{V} = O(n^{-1})$ but it is straightforward to allow for them if their use is worthwhile; for example, it would be an easy matter to employ the location antithetic $\tilde{\psi}^{(i)} = 2\hat{\psi} - \psi^{(i)}$.

There is flexibility in the way the pairs $\psi^{(i)}$, $\eta^{(i)}$ are chosen, depending on the number of antithetics employed and the way the values of ψ and η are combined. For example, one could begin by making a random selection ψ^s of ψ from $\mathrm{N}(\hat{\psi}, \hat{V})$. Next we compute the antithetic value $\tilde{\psi}^s = 2\hat{\psi} - \psi^s$. For each of the values ψ^s and $\tilde{\psi}^s$ one could draw separate values of η from $g(\eta|\psi, Y_n)$, and then employ the two antithetics for each η that are described in Subsection 11.4.3. Thus in the sample there are four values of η combined with each value of ψ so N is a multiple of four and the number of draws of η from the simulation smoother is $N/4$. For estimation of variances due to simulation we need however to note that, since ψ^s and $\tilde{\psi}^s$ are related, there are only $N/8$ independent draws from the joint importance density $g(\psi, \eta|Y_n)$. For the purpose of estimating posterior variances of scalar quantities, assume that $x^*(\eta)$ is a scalar. Then, as in (11.21), the estimate of its posterior variance is

$$\widehat{\mathrm{Var}}[x^*(\eta)|Y_n] = \frac{\sum_{i=1}^N x_i^2 z_i}{\sum_{i=1}^N z_i} - \hat{x}^2. \tag{13.24}$$

Let us now consider the estimation of variance of the estimate $\hat{x}$ of the posterior mean of scalar $x^*(\eta)$ due to simulation. As indicated above the details depend on the way values of ψ and η are combined. For the example we considered, with a single antithetic for ψ and two antithetics for η, combined in the way described, let $\hat{v}_j^\dagger$ be the sum of the eight associated values of $z_i(x_i - \hat{x})$. Then as in (11.23), the estimate of the variance of $\hat{x}$ due to errors of simulation is

$$\widehat{\mathrm{Var}}_s(\hat{x}) = \frac{\sum_{j=1}^{N/8} \hat{v}_j^{\dagger 2}}{\left(\sum_{i=1}^N z_i\right)^2}. \tag{13.25}$$

For the estimation of posterior distribution functions and densities of scalar $x^*(\eta)$, let $I_x(\eta)$ be an indicator which is unity if $x^*(\eta) \leq x$ and is zero if $x^*(\eta) > x$. Then the posterior distribution function is estimated by (11.24) provided that w_i is replaced by z_i. With the same proviso, the posterior density of $x^*(\eta)$ is estimated by (11.25). Samples of independent values from the estimated posterior distribution can be obtained by a method analogous to that described by a method at the end of Subsection 11.5.3.

13.3.3 Posterior analysis of parameter vector

In this section we consider the estimation of posterior means, variances, distribution functions and densities of functions of the parameter vector ψ. Denote by $\nu(\psi)$ the function of ψ whose posterior properties we wish to investigate. Using Bayes theorem, the posterior mean of $\nu(\psi)$ is

$$\begin{aligned}\bar{\nu} &= \mathrm{E}[\nu(\psi)|Y_n]\\ &= \int \nu(\psi)p(\psi|Y_n)\,d\psi\\ &= K\int \nu(\psi)p(\psi)p(Y_n|\psi)\,d\psi\\ &= K\int \nu(\psi)p(\psi)p(\eta, Y_n|\psi)\,d\psi d\eta, \end{aligned} \tag{13.26}$$

where K is a normalising constant. Introducing importance densities $g(\psi|Y_n)$ and $g(\eta|\psi, Y_n)$ as in Subsection 13.3.2, we have

$$\begin{aligned}\bar{\nu} &= K\int \nu(\psi)\frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)}\frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)}g(\psi, \eta|Y_n)\,d\psi d\eta\\ &= K\mathrm{E}_g[\nu(\psi)z^*(\psi, \eta, Y_n)], \end{aligned} \tag{13.27}$$

where E_g denotes expectation with respect to the joint importance density $g(\psi, \eta|Y_n)$ and

$$z^*(\psi, \eta, Y_n) = \frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)}\frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)}.$$

Putting $\nu(\psi) = 1$ in (13.27) we obtain as in (13.20),

$$\bar{\nu} = \frac{\mathrm{E}_g[\nu(\psi)z^*(\psi, \eta, Y_n)]}{\mathrm{E}_g[z^*(\psi, \eta, Y_n)]}. \tag{13.28}$$

In the simulation, take $\psi^{(i)}$ and $\eta^{(i)}$ as in Subsection 13.3.2 and let $\nu_i = \nu(\psi^{(i)})$. Then the estimates $\hat{\nu}$ of $\bar{\nu}$ and $\widehat{\mathrm{Var}}[\nu(\psi)|Y_n]$ of $\mathrm{Var}[\nu(\psi)|Y_n]$ are given by (13.23) and (13.24) by replacing x_i by ν_i. Similarly, the variance of $\hat{\nu}$ due to simulation can, for the antithetics considered in Subsection 11.5.3, be calculated by defining $v_j^\dagger$ as the sum of the eight associated values of $z_i(\nu_i - \bar{\nu})$ and using (13.25) to obtain the estimate $\widehat{\mathrm{Var}}_s(\hat{\nu})$. Estimates of the posterior distribution and density functions are obtained by the indicator function techniques described at the end of Subsection 13.3.2. While $\hat{\nu}$ can be a vector, for the remaining estimates $\nu(\psi)$ has to be a scalar quantity.

The estimate of the posterior density $p[\nu(\psi)|Y_n]$ obtained in this way is essentially a histogram estimate, which is accurate at values of $\nu(\psi)$ near the midpoint of the intervals containing them. An alternative estimate of the posterior density of a particular element of ψ, which is accurate at any value of the element, was proposed by Durbin and Koopman (2000). Without loss of generality take this element to be the first element of ψ and denote it by ψ_1. Denote the remaining elements by ψ_2. Let $g(\psi_2|\psi_1, Y_n)$ be the approximate conditional density of ψ_2 given ψ_1 and Y_n, which is easily obtained by applying standard regression theory to $g(\psi|Y_n)$, where $g(\psi|Y_n) = \mathrm{N}(\hat{\mu}, \hat{V})$. We take $g(\psi_2|\psi_1, Y_n)$ as an importance density in place of $g(\psi|Y_n)$. Then

$$\begin{aligned} p(\psi_1|Y_n) &= \int p(\psi|Y_n)\, d\psi_2 \\ &= K \int p(\psi)p(Y_n|\psi)\, d\psi_2 \\ &= K \int p(\psi)p(\eta, Y_n|\psi)\, d\psi_2 d\eta \\ &= K\mathrm{E}_g[\tilde{z}(\psi, \eta, Y_n)], \end{aligned} \tag{13.29}$$

where E_g denotes expectation with respect to importance density $g(\psi_2|\psi_1, Y_n)$ and

$$\tilde{z}(\psi, \eta, Y_n) = \frac{p(\psi)g(Y_n|\psi)}{g(\psi_2|\psi_1, Y_n)} \frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)}. \tag{13.30}$$

Let $\tilde{\psi}_2^{(i)}$ be a draw from $g(\psi_2|\psi_1, Y_n)$, let $\tilde{\psi}^{(i)} = (\psi_1, \tilde{\psi}_2^{(i)\prime})'$ and let $\tilde{\eta}^{(i)}$ be a draw from $g(\eta|\tilde{\psi}^{(i)}, Y_n)$. Then take

$$\tilde{z}_i = \frac{p\big(\tilde{\psi}^{(i)}\big)g\big(Y_n|\tilde{\psi}^{(i)}\big)}{g\big(\tilde{\psi}_2^{(i)}|\psi_1, Y_n\big)} \frac{p\big(\tilde{\eta}^{(i)}, Y_n|\tilde{\psi}^{(i)}\big)}{g\big(\tilde{\eta}^{(i)}, Y_n|\tilde{\psi}^{(i)}\big)}. \tag{13.31}$$

Now as in (13.28),

$$K^{-1} = \mathrm{E}_g[z^*(\psi, \eta, Y_n)],$$

where E_g denotes expectation with respect to importance density $g(\psi|Y_n)$ $g(\eta|\psi, Y_n)$ and

$$z^*(\psi, \eta, Y_n) = \frac{p(\psi)g(Y_n|\psi)}{g(\psi|Y_n)} \frac{p(\eta, Y_n|\psi)}{g(\eta, Y_n|\psi)}.$$

Let ψ_i^* be a draw from $g(\psi|Y_n)$ and let η_i^* be a draw from $g(\eta|\psi_i^*, Y_n)$. Then take

$$z_i^* = \frac{p(\psi_i^*)g(Y_n|\psi_i^*)}{g(\psi_i^*|Y_n)} \frac{p(\eta_i^*, Y_n|\psi_i^*)}{g(\eta_i^*, Y_n|\psi_i^*)}, \tag{13.32}$$

and estimate $p(\psi_i|Y_n)$ by the simple form

$$\hat{p}(\psi_i|Y_n) = \sum_{i=1}^{N} \tilde{z}_i / \sum_{i=1}^{N} z_i^*. \tag{13.33}$$

The simulations for the numerator and denominator of (13.33) are different since for the numerator only ψ_2 is drawn, whereas for the denominator the whole vector ψ is drawn. The variability of the ratio can be reduced however by employing the same set of $\mathrm{N}(0,1)$ deviates employed for choosing η from $p(\eta|\tilde{\psi}^{(i)}, Y_n)$ in the simulation smoother as for choosing η from $p(\eta|\psi_i^*, Y_n)$. The variability can be reduced further by first selecting ψ_{1i}^* from $g(\psi_1|Y_n)$ and then using the same set of $\mathrm{N}(0,1)$ deviates to select ψ_{2i}^* from $g(\psi_2|\psi_{1i}^*, Y_n)$ as were used to select $\tilde{\psi}_2^{(i)}$ from $g(\psi_2|\psi_1, Y_n)$ when computing $\tilde{z}_i$; in this case $g(\psi^*|Y_n)$ in (13.32) is replaced by $g(\psi_1^*)g(\psi_2^*|\psi_{1i}^*, Y_n)$.

To improve efficiency, antithetics may be used for draws of ψ and η in the way suggested in Subsection 13.3.2.

13.4 Markov chain Monte Carlo methods

An alternative approach to Bayesian analysis based on simulation is provided by the *Markov chain Monte Carlo* (MCMC) method which has received a substantial amount of interest in the statistical and econometric literature on time series. We briefly outline here the basic ideas of MCMC as applied to state space models. Frühwirth-Schnatter (1994) was the first to give a full Bayesian treatments of the linear Gaussian model using MCMC techniques. The proposed algorithms for simulation sample selection were later refined by Carter and Kohn (1994) and de Jong and Shephard (1995). This work resulted in the simulation smoother of Durbin and Koopman (2002) which we discussed in Section 4.9. We showed there how to generate random draws from the conditional densities $p(\varepsilon|Y_n, \psi)$, $p(\eta|Y_n, \psi)$ and $p(\alpha|Y_n, \psi)$ for a given parameter vector ψ. Now we briefly discuss how this technique can be incorporated into a Bayesian MCMC analysis in which we treat the parameter vector as stochastic.

The basic idea is as follows. We evaluate the posterior mean of $x(\alpha)$ or of the parameter vector ψ via simulation by choosing samples from an augmented joint density $p(\psi, \alpha|Y_n)$. In the MCMC procedure, the sampling from this joint density is implemented as a Markov chain. After initialisation for ψ, say $\psi = \psi^{(0)}$ we repeatedly cycle through the two simulation steps:

(1) sample $\alpha^{(i)}$ from $p(\alpha|Y_n, \psi^{(i-1)})$;

(2) sample $\psi^{(i)}$ from $p(\psi|Y_n, \alpha^{(i)})$;

for $i = 1, 2 \ldots$. After a number of 'burning-in' iterations we are allowed to treat the samples from step (2) as being generated from the density $p(\psi|Y_n)$. The attraction of this MCMC scheme is that sampling from conditional densities is easier than sampling from the marginal density $p(\psi|Y_n)$. The circumstances under which subsequent samples from the marginal densities $p(\alpha|Y_n, \psi^{(i-1)})$ and $p(\psi|Y_n, \alpha^{(i)})$ converge to samples from the joint density $p(\psi, \alpha|Y_n)$ are considered in books on MCMC, for example, Gamerman and Lopes (2006). It is not straightforward to develop appropriate diagnostics which indicate whether convergence within the MCMC process has taken place, as is discussed, for example, in Gelman (1995).

There exist various implementations of the basic MCMC algorithm for the state space model. For example, Carlin, Polson and Stoffer (1992) propose samplling individual state vectors from $p(\alpha_t|Y_n, \alpha^t, \psi)$ where α^t is equal to α excluding α_t. It turns out that this approach to sampling is inefficient. It is argued by Frühwirth-Schnatter (1994) that it is more efficient to sample all the state vectors directly from the density $p(\alpha|Y_n, \psi)$. She provides the technical details of implementation. de Jong and Shephard (1995) have developed this approach further by concentrating on the disturbance vectors ε_t and η_t instead of the state vector α_t. The details regarding the resulting simulation smoother were given in Section 4.9.

Implementing the two steps of the MCMC is not as straightforward as suggested so far. Sampling from the density $p(\alpha|Y_n, \psi)$ for a given ψ is done by using the simulation smoother of Section 4.9. Sampling from $p(\psi|Y_n, \alpha)$ depends partly on the model for ψ and is usually only possible up to proportionality. To sample under such circumstances, *accept-reject* algorithms have been developed; for example, the Metropolis algorithm is often used for this purpose. Details and an excellent general review of these matters are given by Gilks, Richardson and Spiegelhalter (1996). Applications to state space models have been developed by Carter and Kohn (1994), Shephard (1994b), Gamerman (1998) and Frühwirth-Schnatter (2006).

In the case of structural time series models of Section 3.2 for which the parameter vector consists only of variances of disturbances associated with the components, the distribution of the parameter vector can be modelled such that sampling from $p(\psi|Y_n, \alpha)$ in step (2) is relatively straightforward. For example, a model for a variance can be based on the inverse gamma distribution with logdensity

$$\log p(\sigma^2|c, s) = -\log \Gamma\left(\frac{c}{2}\right) - \frac{c}{2}\log\frac{s}{2} - \frac{c+2}{2}\log\sigma^2 - \frac{s}{2\sigma^2}, \qquad \text{for } \sigma^2 > 0,$$

and $p(\sigma^2|c, s) = 0$ for $\sigma^2 \leq 0$; see, for example, Poirier (1995). We denote this density by $\sigma^2 \sim \text{IG}(c/2, s/2)$ where c determines the shape and s determines the scale of the distribution. It has the convenient property that if we take this as the prior density of σ^2 and we take a sample $u_1, \ldots, u_n$ of independent $\mathrm{N}(0, \sigma^2)$ variables, the posterior density of σ^2 is

$$p(\sigma^2 | u_1, \ldots, u_n) = \mathrm{IG}\left[(c+n)/2, \left(s + \sum_{i=1}^{n} u_i^2 \right) \bigg/ 2 \right];$$

for further details see, for example, Poirier (1995). For the implementation of step (2) a sample value of σ^2 is chosen from this density. We can take u_t as an element of ε_t or η_t obtained by the simulation smoother in step (1). Further details of this approach are given by Frühwirth-Schnatter (1994, 2006) and Carter and Kohn (1994).

14 Non-Gaussian and nonlinear illustrations

14.1 Introduction

In this chapter we illustrate the methodology of Part II by applying it to different real data sets. In the first example of Section 14.2 we consider the estimation of a multiplicative trend and seasonal model. In Section 14.3 we examine the effects of seat belt legislation on deaths of van drivers due to road accidents in Great Britain modelled by a Poisson distribution. In the third example of Section 14.4 we consider the usefulness of the t-distribution for modelling observation errors in a gas consumption series containing outliers. In Section 14.5 we fit a stochastic volatility model to a series of pound/dollar exchange rates using different methodologies. In the final illustration of Section 14.6 we fit a binary model to the results of the Oxford Cambridge boat race over a long period with many missing observations and we forecast the probability that Oxford will win in 2012.

14.2 Nonlinear decomposition: UK visits abroad

It is common practice in economic time series analyses and seasonal adjustment procedures to take logarithms of the data and to adopt a linear Gaussian time series model for its analysis. The logarithmic transformation converts an exponentially growing trend into a linear trend while it also eliminates or reduces growing seasonal variation and heteroscedasticity in seasonal time series. The logadditive framework appears to work successfully for a model-based decomposition of the time series in trend, seasonal, irregular and other components. It predicates that time series components combine multiplicatively in the implied model for the untransformed series. A full multiplicative model is however not always intended or desired. If heteroscedasticity or changing seasonal variation remains after the logtransformation, applying the logtransformation again is not an attractive solution. Also, if the data is already supplied in units measuring proportional changes, applying the logtransformation can complicate model interpretation.

Koopman and Lee (2009) propose a nonlinear unobserved component time series model that can be used when the time series, possibly after a logtransformation, is not appropriate for an analysis based on the linear model. They generalise the basic structural model of Subsection 3.2.3 by scaling the amplitude of the seasonal component γ_t via an exponential transformation of the trend component μ_t. The observed time series y_t, either in levels or in logs, is decomposed by the nonlinear model

$$y_t = \mu_t + \exp(c_0 + c_\mu \mu_t)\gamma_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma^2_\varepsilon), \qquad t = 1, \ldots, n, \qquad (14.1)$$

where c_0 and c_μ are an unknown fixed coefficients while dynamic specifications of the trend component μ_t are discussed in Subsection 3.2.1 and those of the seasonal component γ_t in Subsection 3.2.2. Coefficient c_0 scales the seasonal effect and therefore we restrict $\sigma^2_\omega = 1$ in the seasonal models of Subsection 3.2.2. The sign of the coefficient c_μ determines whether the seasonal variation increases or decreases when a positive change in the trend occurs. The model reduces to the linear specification of Subsection 3.2.3 when c_μ is zero. The overall time-varying amplitude of the seasonal component is determined by the combined effect $c_0 + c_\mu \mu_t$.

We consider a data set of monthly visits abroad by UK residents from January 1980 to December 2006. The data is compiled by the Office for National Statistics (ONS), based on the International Passenger Survey. Fig. 14.1 presents the time series in levels and in logs. The time series of visits abroad shows a clear upwards trend, a pronounced seasonal pattern, and a steady increase of the seasonal variation over time. After applying the logtransformation, the increase of

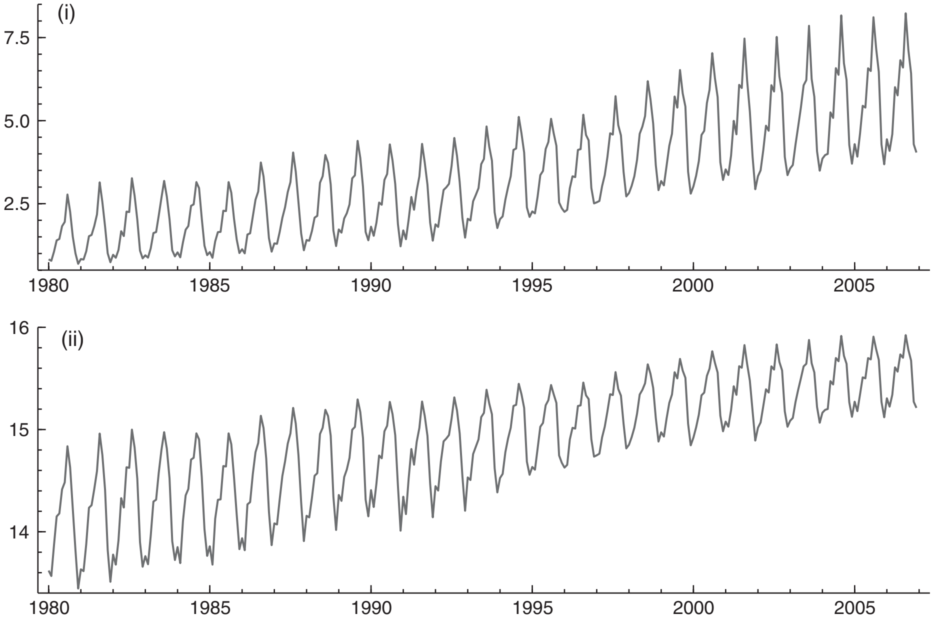

Fig. 14.1 Visits of UK residents abroad (i) in levels (million); (ii) in logarithms.

seasonal variation has been converted into a decrease. This may indicate that the logtransformation is not particularly appropriate for this series.

We therefore consider the model given by (14.1) with a cycle component ψ_t discussed in Subsection 3.2.4 added to capture economic business cycle behaviour from the data, that is $y_t = \mu_t + \psi_t + \exp(c_0 + c_\mu \mu_t)\gamma_t + \varepsilon_t$ with irregular $\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. The trend μ_t is specified as the local linear trend (3.2) with $\sigma_\xi^2 = 0$ (smooth trend), the seasonal γ_t has the trigonometric specification (3.5) and the cycle ψ_t is specified as c_t in (3.13). Due to the seasonal effect $\exp(c_0+c_\mu\mu_t)\gamma_t$, we obtain a nonlinear observation equation $y_t = Z(\alpha_t)+\varepsilon_t$ where the state vector α_t consists of elements associated with the trend, seasonal and cycle components; see Subsections 3.2.3 and 3.2.4. The model is a special case of the nonlinear state space model discussed in Section 9.7.

We apply the extended Kalman filter as developed in Section 10.2 to our model. The Gaussian likelihood function (7.2), with v_t and F_t computed by the extended Kalman filter (10.4) is treated as an approximated likelihood function. The initialisation of the extended Kalman filter is discussed in Koopman and Lee (2009). The numerical optimisation of the approximate loglikelihood function produces the parameter estimates given by

$$\begin{aligned} \hat\sigma_\varepsilon = 0.116, \quad \hat\sigma_\zeta = 0.00090, \quad \hat c_0 = -5.098, \quad \hat c_\mu = 0.0984,& \\ \hat\sigma_\kappa = 0.00088, \quad \hat\rho = 0.921, \quad 2\pi/\hat{\lambda^c} = 589.& \end{aligned} \tag{14.2}$$

The seasonal component γ_t is scaled by $\exp(c_0 + c_\mu\mu_t)$. In Fig. 14.2, panel (i) presents the scaled seasonal component $\exp(c_0 + c_\mu\mu_t)\gamma_t$ and the unscaled component γ_t as estimated by the extended Kalman smoother discussed in Subsection 10.4.1. The scaled component is changing largely due to the trend component which is plotted in panel (ii) of Fig. 14.2. The unscaled component shows a more stable pattern with almost a constant amplitude over time. A more detailed discussion of this analysis is given by Koopman and Lee (2009).

14.3 Poisson density: van drivers killed in Great Britain

The assessment for the Department of Transport of the effects of seat belt legislation on road traffic accidents in Great Britain, described by Harvey and Durbin (1986) and also discussed in Section 8.2, was based on linear Gaussian methods as described in Part I. One series that was excluded from this study was the monthly numbers of light goods vehicle (van) drivers killed in road accidents from 1969 to 1984. The numbers of deaths of van drivers were too small to justify the use of the linear Gaussian model. A better model for the data is based on the Poisson distribution with mean $\exp(\theta_t)$ and density

$$p(y_t|\theta_t) = \exp\{\theta_t' y_t - \exp(\theta_t) - \log y_t!\}, \qquad t = 1, \ldots, n, \tag{14.3}$$

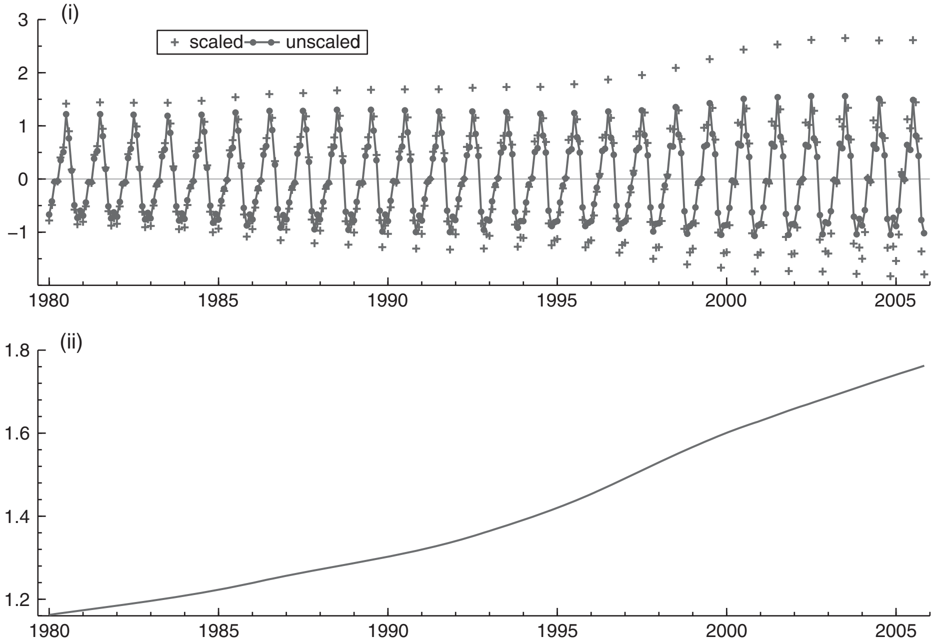

Fig. 14.2 Visits of UK residents abroad: (i) smooth estimates of the scaled and unscaled seasonal components obtained by the extended Kalman filter and smoother; (ii) scaling process $\exp(c_0 + c_\mu \mu_t)$ with μ_t replaced by its smoothed estimate.

as discussed in Subsection 9.3.1. We model θ_t by the relation

$$\theta_t = \mu_t + \gamma_t + \lambda x_t,$$

where the trend μ_t is the random walk

$$\mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \sim \mathrm{N}\big(0, \sigma_\eta^2\big), \tag{14.4}$$

λ is the intervention parameter which measures the effects of the seat belt law, x_t is an indicator variable for the post legislation period and the monthly seasonal γ_t is generated by

$$\sum_{j=0}^{11} \gamma_{t+1-j} = \omega_t, \qquad \omega_t \sim \mathrm{N}\big(0, \sigma_\omega^2\big). \tag{14.5}$$

The disturbances η_t and ω_t are mutually independent Gaussian white noise terms with variances $\sigma_\eta^2 = \exp(\psi_\eta)$ and $\sigma_\omega^2 = \exp(\psi_\omega)$, respectively. The parameter estimates are reported by Durbin and Koopman (1997) as $\hat{\sigma}_\eta = \exp(\hat{\psi}_\eta) = \exp(-3.708) = 0.0245$ and $\hat{\sigma}_\omega = 0$. The fact that $\hat{\sigma}_\omega = 0$ implies that the seasonal is constant over time.

For the Poisson model we have $b_t(\theta_t) = \exp(\theta_t)$. As in Subsection 10.6.4 we have $\dot{b}_t = \ddot{b}_t = \exp(\tilde{\theta}_t)$, so we take

$$A_t = \exp(-\tilde{\theta}_t), \qquad x_t = \tilde{\theta}_t + A_t y_t - 1,$$

where $\tilde{\theta}_t$ is some trial value for θ_t with $t = 1, \ldots, n$. The iterative process for determining the approximating model as described in Section 10.6 converges quickly; usually, between three and five iterations are needed for the Poisson model. The estimated signal $\mu_t + \lambda x_t$ for ψ_η fixed at $\hat{\psi}_\eta$ is computed and its exponentiated values are plotted together with the raw data in panel (i) of Fig. 14.3.

The main objective of the analysis is the estimation of the effect of the seat belt law on the number of deaths. Here, this is measured by λ which is estimated as the value -0.278 with standard error 0.114. The estimate of λ corresponds to a reduction in the number of deaths of 24%. It is clear from the standard

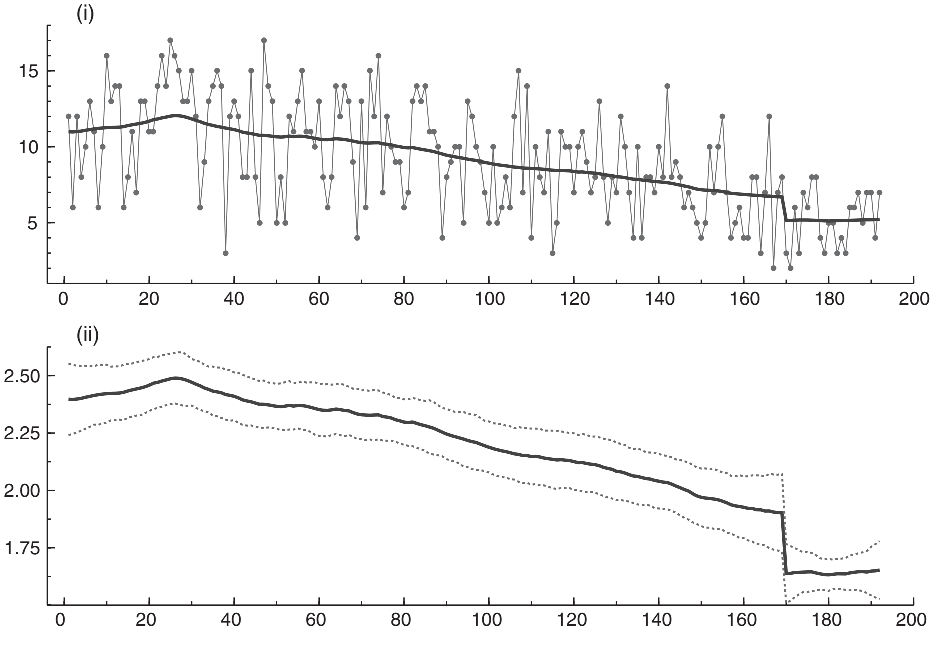

Fig. 14.3 Numbers of van drivers killed: (i) observed time series counts and estimated level including intervention, $\exp(\mu_t + \lambda x_t)$; (ii) estimated level including intervention, $\mu_t + \lambda x_t$, and its confidence interval based on two times standard error.

error that the seat belt law has given some significant reduction to the number of deaths; this is confirmed visually in panel (ii) of Fig. 14.3. What we learn from this exercise so far as the underlying real investigation is concerned is that up to the point where the law was introduced there was a slow regular decline in the number of deaths coupled with a constant multiplicative seasonal pattern, while at that point there was an abrupt drop in the trend of around 25%; afterwards, the trend appeared to flatten out, with the seasonal pattern remaining the same. A more detailed analysis on the basis of this model, and including a Bayesian analysis, is presented by Durbin and Koopman (2000).

14.4 Heavy-tailed density: outlier in gas consumption

In this example we analyse the logged quarterly demand for gas in the UK from 1960 to 1986 which is a series from the standard data set provided by Koopman, Harvey, Doornik and Shephard (2010). We use a structural time series model of the basic form as discussed in Section 3.2:

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \tag{14.6}$$

where μ_t is the local linear trend, γ_t is the seasonal and ε_t is the observation disturbance. The purpose of the investigation underlying the analysis is to study the seasonal pattern in the data with a view to seasonally adjusting the series. It is known that for most of the series the seasonal component changes smoothly over time, but it is also known that there was a disruption in the gas supply in the third and fourth quarters of 1970 which leads to a distortion in the seasonal pattern when a standard analysis based on a Gaussian density for ε_t is employed. The question under investigation is whether the use of a heavy-tailed density for ε_t would improve the estimation of the seasonal in 1970.

To model ε_t we use the t-distribution as in Subsection 9.4.1 with logdensity

$$\log p(\varepsilon_t) = \log a\,(\nu) + \frac{1}{2}\log\lambda - \frac{\nu+1}{2}\log\left(1+\lambda\varepsilon_t^2\right), \tag{14.7}$$

where

$$a(\nu) = \frac{\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \qquad \lambda^{-1} = (\nu-2)\,\sigma_\varepsilon^2, \qquad \nu > 2, \qquad t = 1,\dots,n.$$

The mean of ε_t is zero and the variance is σ_ε^2 for any ν degrees of freedom which need not be an integer. The approximating model can be obtained by the methods described in Section 10.8. When we use the first derivative only, we obtain

$$A_t = \frac{1}{\nu+1}\tilde{\varepsilon}_t^2 + \frac{\nu-2}{\nu+1}\sigma_\varepsilon^2,$$

The iterative scheme is started with $A_t = \sigma_\varepsilon^2$, for $t = 1, \ldots, n$. The number of iterations required for a reasonable level of convergence using the t-distribution is usually higher than for densities from the exponential family; for this example we required around ten iterations. In the classical analysis, the parameters of the model, including the degrees of freedom ν, were estimated by Monte Carlo maximum likelihood as described in Subsection 11.6.2; the estimated value for ν was 12.8.

We now compare the estimated seasonal and irregular components based on the Gaussian model and the model with a t-distribution for ε_t. Fig. 14.4 provide the graphs of the estimated seasonal and irregular for the Gaussian model and the t-model. The most striking feature of these graphs is the greater effectiveness with which the t-model picks and corrects for the outlier relative to the Gaussian model. The t-model estimates are based on 250 simulation samples from the simulation smoother with four antithetics for each sample. We learn from the analysis that the change over time of the seasonal pattern in the data is in fact smooth. We also learn that if model (14.6) is to be used to estimate the seasonal for this or similar cases with outliers in the observations, then a Gaussian model for ε_t is inappropriate and a heavy-tailed model should be used.

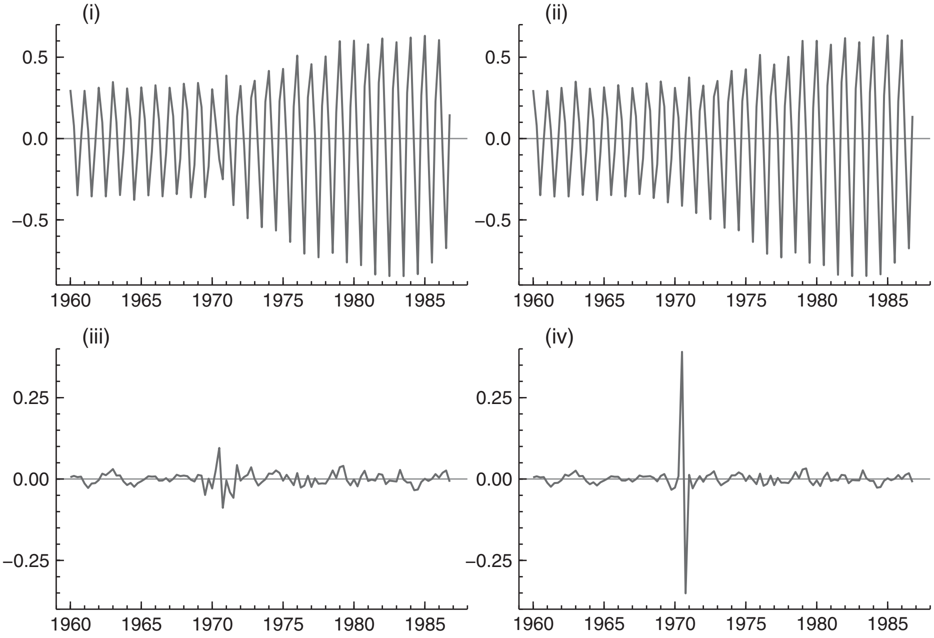

Fig. 14.4 Analyses of gas data: (i) estimated seasonal component from Gaussian model; (ii) estimated seasonal component from t model; (iii) estimated irregular from Gaussian model; (iv) estimated irregular from t model.

14.5 Volatility: pound/dollar daily exchange rates

The stochastic volatility (SV) model is discussed in detail in Section 9.5. In our empirical illustration for the pound/dollar daily exchange rates we consider a basic version of the SV model. The time series of exchange rates is from 1/10/81 to 28/6/85 and have been used by Harvey, Ruiz and Shephard (1994). Denoting the daily exchange rate by x_t, the observations we consider are given by $y_t = \log x_t, x_{t-1}$ for $t = 1, \dots, n$. A zero-mean stochastic volatility model of the form

$$
\begin{aligned}
y_t &= \sigma \exp\left(\frac{1}{2}\theta_t\right) u_t, \qquad && u_t \sim \mathrm{N}(0,1), \qquad && t = 1, \dots, n, \\
\theta_{t+1} &= \phi\theta_t + \eta_t, && \eta_t \sim \mathrm{N}\big(0, \sigma_\eta^2\big), && 0 < \phi < 1,
\end{aligned}
\tag{14.8}
$$

was used for analysing these data by Harvey, Ruiz and Shephard (1994). The purpose of the investigations for which this type of analysis is carried out is to study the structure of the volatility of price ratios in the market, which is of considerable interest to financial analysts. The level of θ_t determines the amount of volatility and the value of ϕ measures the autocorrelation present in the volatility process.

14.5.1 Data transformation analysis

We start with providing an approximate solution based on the linear model as suggested in Section 10.5. After the observations y_t are transformed into $\log y_t^2$, we consider the linear model (10.32), that is

$$
\log y_t^2 = \kappa + \theta_t + \xi_t, \qquad t = 1, \dots, n, \tag{14.9}
$$

where κ is an unknown constant and ξ_t is a mean-zero disturbance which is not normally distributed. Given the model is linear, we can proceed approximately with the methods developed in Part I. This approach is taken by Harvey, Ruiz and Shephard (1994) who refer to it as a quasi-maximum likelihood (QML) method. Parameter estimation is done via the Kalman filter; smoothed estimates of the volatility component, θ_t, are constructed and forecasts of volatility can be generated. One of the attractions of the QML approach is that it can be carried out straightforwardly using standard software such as *STAMP* of Koopman, Harvey, Doornik and Shephard (2010). This is an advantage compared to the more involved simulation-based methods.

In our illustration of the QML method, we use the same data as analysed by Harvey, Ruiz and Shephard (1994) in which y_t is the first difference of the logged exchange rate between pound sterling and US dollar. To avoid taking logs of zero values, it is common practice to take deviations from its sample mean. The resulting mean-corrected logreturns are then analysed by the QML method.

Parameter estimation is carried out carried out quickly by the Kalman filter. We obtain the following QML estimates:

$$\begin{array}{rclrclrcl}
\hat{\sigma}_\xi &=& 2.1521, & \hat{\psi}_1 &=& \log\hat{\sigma}_\xi \;=\; 0.7665, & \mathrm{SE}(\hat{\psi}_1) &=& 0.0236,\\
\hat{\sigma}_\eta &=& 0.8035, & \hat{\psi}_2 &=& \log\hat{\sigma}_\eta \;=\; -0.2188, & \mathrm{SE}(\hat{\psi}_2) &=& 0.5702,\\
\hat{\phi} &=& 0.9950, & \hat{\psi}_3 &=& \log\frac{\hat{\phi}}{1-\hat{\phi}} \;=\; 5.3005, & \mathrm{SE}(\hat{\psi}_3) &=& 1.6245,
\end{array}$$

where SE denotes the standard error of the maximum likelihood estimator. We present the results in this form since we estimate the logtransformed parameters, so the standard errors that we calculate apply to them and not to the original parameters of interest.

Once the parameters are estimated, we can compute the smoothed estimate of the signal θ_t using standard Kalman filter and smoother methods. The logreturns for the pound/US dollar exhange series (adjusted for the mean, that is y_t) is depicted in panel (i) of Fig. 14.5. The signal extraction results are presented in panels (ii) and (iii) of Fig. 14.5. In panel (ii) we present the transformed data $\log y_t^2$ together with the smoothed estimate of θ_t from the Kalman filter

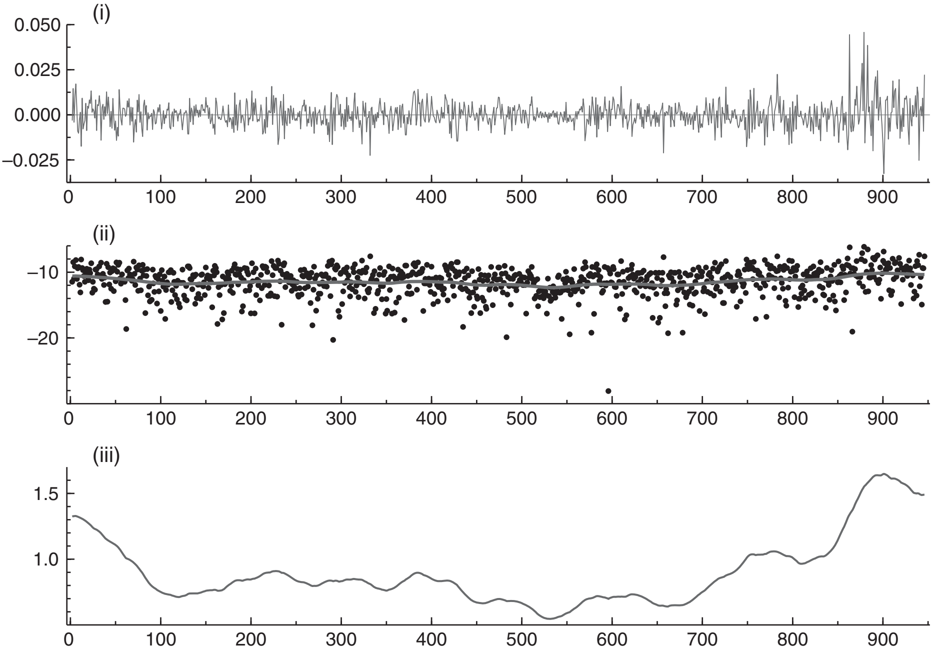

Fig. 14.5 Analyses of pound–dollar exchange rates: (i) daily logreturns of exchange rates, mean-corrected, denoted as y_t; (ii) the $\log y_t^2$ series with the smoothed estimate of $\kappa + \theta_t$;(ii) smoothed estimate of volatility measure $\exp(\theta_t\,/\,2)$.

and smoother using the linear model (10.32). The smoothed estimate of the volatility, which we measure as $\exp(\theta_t/2)$, is displayed in panel (iii).

14.5.2 Estimation via importance sampling

To illustrate the maximum likelihood estimation of parameters in the SV model using importance sampling, we consider the Gaussian logdensity of model (14.8) which is given by

$$\log p(y_t|\theta_t) = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\theta_t - \frac{y_t^2}{2\sigma^2}\exp(-\theta_t). \tag{14.10}$$

The linear approximating model based on the estimated mode of θ_t can be obtained by the method of Subsection 10.6.5 with

$$A_t = 2\sigma^2\frac{\exp(\tilde{\theta}_t)}{y_t^2}, \qquad x_t = \tilde{\theta}_t - \frac{1}{2}\tilde{H}_t + 1,$$

for which A_t is always positive. The iterative process can be started with $A_t = 2$ and $x_t = \log(y_t^2/\sigma^2)$, for $t = 1, \ldots, n$, since it follows from (14.8) that $y_t^2/\sigma^2 \approx \exp(\theta_t)$. When y_t is zero or very close to zero, it should be replaced by a small constant value to avoid numerical problems; this device is only needed to obtain the approximating model so we do not depart from our exact treatment. The number of iterations required is usually fewer than ten.

Firstly we focus on the estimation of the parameters. For the method of importance sampling, we take $N = 100$ for computing the loglikelihood function. We carry out the computations as detailed in Section 11.6. During the estimation process, we take the same random numbers for each loglikelihood evaluation so that the loglikelihood is a smooth function of the parameters. We then obtain the following estimates, after convergence of the numerical optimisation:

$$\begin{aligned}
\hat{\sigma} &= 0.6338, & \hat{\psi}_1 &= \log\hat{\sigma} = -0.4561, & \mathrm{SE}(\hat{\psi}_1) &= 0.1033,\\
\hat{\sigma}_\eta &= 0.1726, & \hat{\psi}_2 &= \log\hat{\sigma}_\eta = -1.7569, & \mathrm{SE}(\hat{\psi}_2) &= 0.2170,\\
\hat{\phi} &= 0.9731, & \hat{\psi}_3 &= \log\frac{\hat{\phi}}{1-\hat{\phi}} = 3.5876, & \mathrm{SE}(\hat{\psi}_3) &= 0.5007,
\end{aligned}$$

where SE denotes the standard error of the maximum likelihood estimator which applies to the logtransformed parameters and not to the original parameters of interest.

Secondly our aim is to estimate the underlying volatility θ_t via importance sampling. In panel (i) of Fig. 14.6 we present the data as absolute values of the first differences together with the smoothed estimate of the volatility component θ_t. We observe that the estimates capture the volatility features in the time

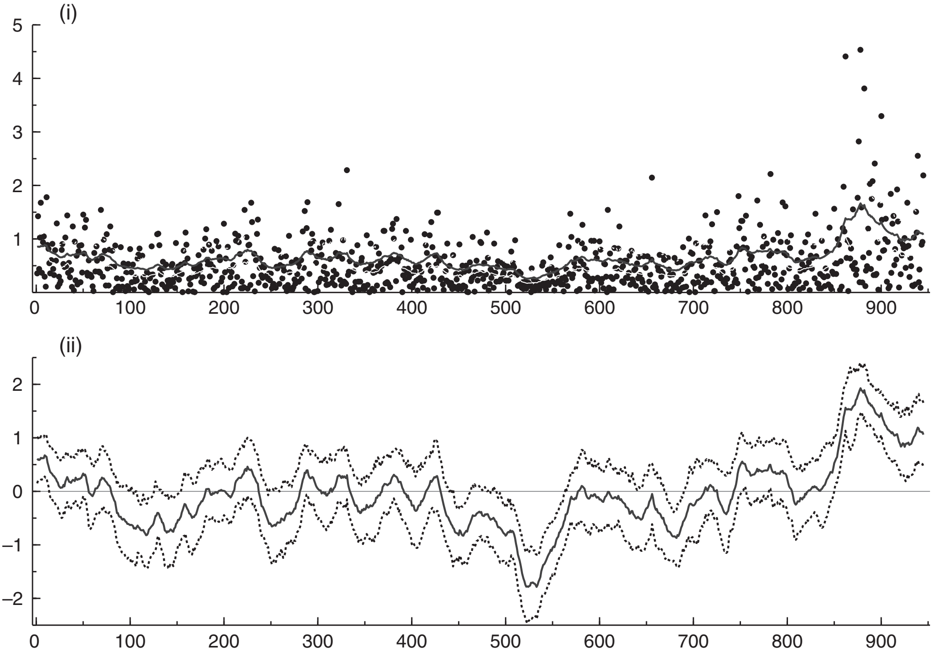

Fig. 14.6 Analyses of pound–dollar exchange rates: (i) difference in absolute values (dots) and smoothed estimate of θ_t; (ii) smoothed estimate of θ_t with 90% confidence interval.

series accurately. Panel (ii) presents the same volatility estimates but here with their 90% confidence interval which are based on standard errors and are also computed by the importance sampling method of Section 11.4.

14.5.3 Particle filtering illustration

To estimate the volatility by filtering for the pound exchange series, we consider both the bootstrap filter and the auxiliary filter and we assess their accuracy in relation to each other. The bootstrap filter is discussed in Subsection 12.4.2 and consists of the following steps for the SV model at a fixed time t and for a given set of particles $\theta_{t-1}^{(1)}, \ldots, \theta_{t-1}^{(N)}$:

(i) Draw N values $\tilde{\theta}_t^{(i)} \sim \mathrm{N}(\phi\theta_{t-1}^{(i)}, \sigma_\eta^2)$.

(ii) Compute the corresponding weights $\tilde{w}_t^{(i)}$

$$\tilde{w}_t^{(i)} = \exp\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\theta_t^{(i)} - \frac{1}{2\sigma^2}\exp(-\theta_t^{(i)})y_t^2\right), \qquad i = 1, \ldots, N,$$

and normalise the weights to obtain $w_t^{(i)}$.

(iii) Compute

$$\hat{a}_{t|t} = \sum_{i=1}^{N} w_t^{(i)} \tilde{\theta}_t^{(i)}, \qquad \hat{P}_{t|t} = \sum_{i=1}^{N} w_t^{(i)} \tilde{\theta}_t^{(i)\,2} - \hat{a}_{t|t}^2.$$

(iv) Select N new independent particles $\alpha_t^{(i)}$ via stratified sampling; see Subsection 12.3.3.

The implementation of the bootstrap filter is straightforward. To monitor its performance, we compute the efficient sample size (ESS) variable for each t. In panel (i) of Fig. 14.7 we present the ESS variable for each t. In many occasions the number of active particles is sufficiently high. However, at various points in time, the bootstrap filter detoriates and the number of effective particles is below 7500. The filtered estimate $\hat{a}_{t|t}$ of logvolatility θ_t is displayed in panel (ii) together with its 90% confidence interval based on $\hat{P}_{t|t}$. Although the volatility changes over time have the same pattern as the smoothed estimate of logvolatility displayed in panel (ii) of Fig. 14.6, the filtered estimate exhibits a more noisier estimate of logvolatility.

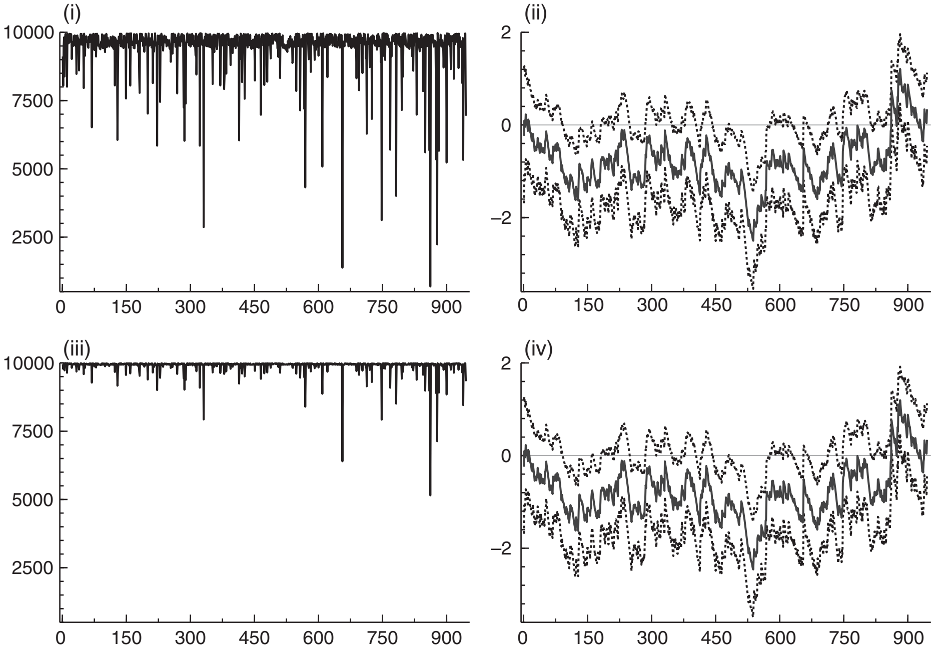

Fig. 14.7 Analyses of pound–dollar exchange rates using bootstrap and auxiliary filters: (i) effective sample size ESS for bootstrap filter, (ii) filtered estimate of θ_t with 90% confidence interval obtained from bootstrap filter, (iii) as (i) for auxiliary filter, (iv) as (ii) obtained from auxiliary filter.

The auxiliary filter is discussed in Section 12.5 and is implemented for the basic SV model. The implementation for $N = 10,000$ is more involved when compared to the bootstrap filter. In panel (iii) of Fig. 14.7 we present the ESS variable for the auxiliary filter; it indicates that the number of effective samples is higher compared to the bootstrap filter. For most t, the number of effective particles is close to $N = 10,000$. The filtered estimate $\hat{a}_{t|t}$ of logvolatility θ_t is displayed in panel (iv) together with its 90% confidence interval based on $\hat{P}_{t|t}$. The filtered estimates from the bootstrap and auxiliary particle filters are virtually the same.

14.6 Binary density: Oxford–Cambridge boat race

In the last illustration we consider the outcomes of the annual boat race between teams representing the universities of Oxford and Cambridge. The race takes place from Putney to Mortlake on the River Thames in the month of March or April. The first took place in 1829 and was won by Oxford and, at the time of writing, the last took place in 2011 and was also won by Oxford; in the year 2010 Cambridge won in an epic battle and denied Oxford the hat-trick. In the first edition of the book, which we finished writing in 2000, we forecasted the probability for a Cambridge win in 2001 as 0.67 and, indeed, Cambridge did win in 2001.

There have been some occasions, especially in the nineteenth century, when the race took place elsewhere and in other months. In the years of both World Wars the race did not take place and there were also some years when the race finished with a dead heat or some other irregularity took place. Thus the time series of yearly outcomes contains missing observations for the years: 1830–1835, 1837, 1838, 1843, 1844, 1847, 1848, 1850, 1851, 1853, 1855, 1877, 1915–1919 and 1940–1945. In Fig. 14.8 the positions of the missing values are displayed as black dots with value 0.5. However, in the analysis we deal with these missing observations as described in Subsection 11.5.5.

The appropriate model for the boat race data is the binary distribution as described in Subsection 9.3.2. We take $y_t = 1$ if Cambridge wins and $y_t = 0$ if Oxford wins. Denoting the probability that Cambridge wins in year t by π_t, as in Subsection 9.3.2 we take $\theta_t = \log[\pi_t/(1-\pi_t)]$. A winner this year is likely to be a winner next year because of overlapping crew membership, training methods and other factors. We model the transformed probability by the random walk

$$\theta_{t+1} = \theta_t + \eta_t, \qquad \eta_t \sim \mathrm{N}\big(0, \sigma_\eta^2\big),$$

where η_t is serially uncorrelated for $t = 1, \ldots, n$.

The method based on mode estimation described in Subsection 10.6.4 provides the approximating model for this case and maximum likelihood estimation for the unknown variance σ_η^2 is carried out as described in Subsection 11.6.

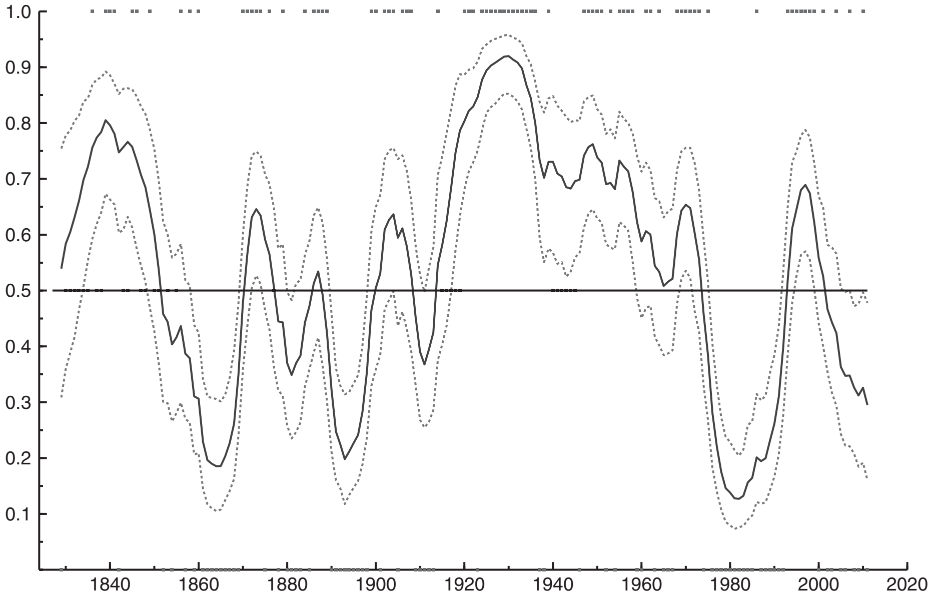

Fig. 14.8 Dot at zero is a win for Oxford, dot at one is a win for Cambridge and dot at 0.5 is a missing value; the solid line is the probability of a Cambridge win and the dotted lines constitute the 50% (asymmetric) confidence interval.

We have estimated the variance as $\hat{\sigma}_\eta^2 = 0.330$. The estimated mean of the probability π_t, indicating a win for Cambridge in year t, is computed using the method described in Subsection 11.5. The resulting time series of π_t is given in Fig. 14.8. The forecasted probability for a Cambridge win in 2012 is 0.30 and therefore we expect the Oxford team to win.

References

Aguilar, O. and M. West (2000). Bayesian dynamic factor models and portfolio allocation. *J. Business and Economic Statist. 18*, 338–57.

Akaike, H. and G. Kitagawa (Eds.) (1999). *The Practice of Time Series Analysis.* New York: Springer-Verlag.

Andersen, T., T. Bollerslev, F. Diebold, and P. Labys (2003). Modelling and forecasting realized volatility. *Econometrica 71*, 529–626.

Anderson, B. D. O. and J. B. Moore (1979). *Optimal Filtering.* Englewood Cliffs: Prentice-Hall [Reprinted by Dover in 2005].

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). New York: John Wiley & Sons.

Ansley, C. F. and R. Kohn (1985). Estimation, filtering and smoothing in state space models with incompletely specified initial conditions. *Annals of Statistics 13*, 1286–1316.

Ansley, C. F. and R. Kohn (1986). Prediction mean square error for state space models with estimated parameters. *Biometrika 73*, 467–74.

Arulampalam, M. S., S. Maskell, N. J. Gordon, and T. Clapp (2002). A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing 50*, 174–88.

Asai, M. and M. McAleer (2005). Multivariate stochastic volatility. Technical report, Tokyo Metropolitan University.

Åström, K. J. (1970). *Introduction to Stochastic Control Theory.* New York: Academic Press [Reprinted by Dover in 2006].

Atkinson, A. C. (1985). *Plots, Transformations and Regression.* Oxford: Clarendon Press.

Balke, N. S. (1993). Detecting level shifts in time series. *J. Business and Economic Statist. 11*, 81–92.

Barndorff-Nielsen, O. E. and N. Shephard (2001). Non-Gaussian OU based models and some of their uses in financial economics (with discussion). *J. Royal Statistical Society B 63*, 167–241.

Basawa, I. V., V. P. Godambe, and R. L. Taylor (Eds.) (1997). *Selected Proceedings of Athens, Georgia Symposium on Estimating Functions.* Hayward, California: Institute of Mathematical Statistics.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory.* Chichester: John Wiley.

Bijleveld, F., J. Commandeur, P. Gould, and S. J. Koopman (2008). Model-based measurement of latent risk in time series with applications. *J. Royal Statistical Society A 171*, 265–77.

Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the Business and Economic Statistics Section*, 177–81.

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *J. Econometrics 51*, 307–327.

Bollerslev, T., R. F. Engle, and D. B. Nelson (1994). ARCH Models. In R. F. Engle and D. McFadden (Eds.), *The Handbook of Econometrics, Volume 4*, pp. 2959–3038. Amsterdam: North-Holland.

Bowman, K. O. and L. R. Shenton (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika 62*, 243–50.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis, Forecasting and Control* (3rd ed.). San Francisco: Holden-Day.

Box, G. E. P. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.

Box, G. E. P. and G. C. Tiao (1975). Intervention analysis with applications to economic and environmental problems. *J. American Statistical Association 70*, 70–79.

Breidt, F. J., N. Crato, and P. de Lima (1998). On the detection and estimation of long memory in stochastic volatility. *J. Econometrics 83*, 325–48.

Brockwell, A. E. (2007). Likelihood-based analysis of a class of generalized long-memory time series models. *J. Time Series Analysis 28*, 386–407.

Brockwell, P. J. and R. A. Davis (1987). *Time Series: Theory and Methods.* New York: Springer-Verlag.

Bryson, A. E. and Y. C. Ho (1969). *Applied Optimal Control.* Massachusetts: Blaisdell.

Burman, J. P. (1980). Seasonal adjustment by signal extraction. *J. Royal Statistical Society A 143*, 321–37.

Burns, A. and W. Mitchell (1946). Measuring Business cycles. Working paper, NBER, New York.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1997). *The Econometrics of Financial Markets.* Princeton, New Jersey: Princeton University Press.

Cappé, O., E. Moulines, and T. Rydén (2005). *Inference in Hidden Markov Models.* New York: Springer.

Cargnoni, C., P. Muller, and M. West (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *J. American Statistical Association 92*, 640–47.

Carlin, B. P., N. G. Polson, and D. S. Stoffer (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modelling. *J. American Statistical Association 87*, 493–500.

Carpenter, J. R., P. Clifford, and P. Fearnhead (1999). An improved particle filter for non-linear problems. *IEE Proceedings. Part F: Radar and Sonar Navigation 146*, 2–7.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*, 541–53.

Carter, C. K. and R. Kohn (1996). Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika 83*, 589–601.

Carter, C. K. and R. Kohn (1997). Semiparameteric Bayesian inference for time series with mixed spectra. *J. Royal Statistical Society B 59*, 255–68.

Chatfield, C. (2003). *The Analysis of Time Series: An Introduction* (6th ed.). London: Chapman & Hall.

Chib, S., F. Nardari, and N. Shephard (2006). Analysis of high dimensional multivariate stochastic volatility models. *J. Econometrics*, *134*, 341–71.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo and its applications to Bayesian inference. *Annals of Statistics 32*, 2385–2411.

Chu-Chun-Lin, S. and P. de Jong (1993). A note on fast smoothing. Discussion paper, University of British Columbia.

Cobb, G. W. (1978). The problem of the Nile: conditional solution to a change point problem. *Biometrika 65*, 243–51.

Commandeur, J. and S. J. Koopman (2007). *An Introduction to State Space Time Series Analysis.* Oxford: Oxford University Press.

Commandeur, J., S. J. Koopman, and M. Ooms (2011). Statistical software for state space methods. *Journal of Statistical Software 41*, Issue 1.

Cook, R. D. and S. Weisberg (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

Creal, D. D. (2012). A survey of sequential Monte Carlo methods for economics and finance. *Econometric Reviews 31*, 245–96.

Danielsson, J. and J. F. Richard (1993). Accelerated Gaussian importance sampler with application to dynamic latent variable models. *J. Applied Econometrics 8*, S153–S174.

Davidson, R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics.* Oxford: Oxford University Press.

de Jong, P. (1988a). A cross validation filter for time series models. *Biometrika 75*, 594–600.

de Jong, P. (1988b). The likelihood for a state space model. *Biometrika 75*, 165–69.

de Jong, P. (1989). Smoothing and interpolation with the state space model. *J. American Statistical Association 84*, 1085–8.

de Jong, P. (1991). The diffuse Kalman filter. *Annals of Statistics 19*, 1073–83.

de Jong, P. (1998). Fixed interval smoothing. Discussion paper, London School of Economics.

de Jong, P. and M. J. MacKinnon (1988). Covariances for smoothed estimates in state space models. *Biometrika 75*, 601–2.

de Jong, P. and J. Penzer (1998). Diagnosing shocks in time series. *J. American Statistical Association 93*, 796–806.

de Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika 82*, 339–50.

Diebold, F. and C. Li (2006). Forecasting the term structure of government bond yields. *J. Econometrics 130*, 337–64.

Diebold, F., S. Rudebusch, and S. Aruoba (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. *J. Econometrics 131*, 309–338.

Doornik, J. A. (2010). *Object-Oriented Matrix Programming using Ox 6.0.* London: Timberlake Consultants Ltd. See http://www.doornik.com [accessed 20 October 2011].

Doran, H. E. (1992). Constraining Kalman filter and smoothing estimates to satisfy time-varying restrictions. *Rev. Economics and Statistics 74*, 568–72.

Doucet, A., N. de Freitas, and N. J. Gordon (2001). *Sequential Monte Carlo Methods in Practice.* New York: Springer Verlag.

Doucet, A., S. J. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing 10*(3), 197–208.

Doz, C., D. Giannone, and L. Reichlin (2012). A quasi maximum likelihood approach for large approximate dynamic factor models. *Rev. Economics and Statistics*, forthcoming.

Duncan, D. B. and S. D. Horn (1972). Linear dynamic regression from the viewpoint of regression analysis. *J. American Statistical Association 67*, 815–21.

Durbin, J. (1960). Estimation of parameters in time series regression models. *J. Royal Statistical Society B 22*, 139–53.

Durbin, J. (1997). Optimal estimating equations for state vectors in non-Gaussian and nonlinear estimating equations. In I. V. Basawa, V. P. Godambe and R. L. Taylor (Eds.), *Selected Proceedings of Athens, Georgia Symposium on Estimating Functions*, Hayward California: Institute of Mathematical Statistics.

Durbin, J. (2000a). Contribution to discussion of Harvey and Chung (2000). *J. Royal Statistical Society A 163*, 303–39.

Durbin, J. (2000b). The state space approach to time series analysis and its potential for official statistics, (The Foreman lecture). *Australian and New Zealand J. of Statistics 42*, 1–23.

Durbin, J. and A. C. Harvey (1985). The effects of seat belt legislation on road casualties in Great Britain: report on assessment of statistical evidence. Annexe to Compulsory Seat Belt Wearing Report, Department of Transport, London, HMSO.

Durbin, J. and S. J. Koopman (1992). Filtering, smoothing and estimation for time series models when the observations come from exponential family distributions. Unpublished paper: Department of Statistics, LSE.

Durbin, J. and S. J. Koopman (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika 84*, 669–84.

Durbin, J. and S. J. Koopman (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *J. Royal Statistical Society B 62*, 3–56.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89*, 603–16.

Durbin, J. and B. Quenneville (1997). Benchmarking by state space models. *International Statistical Review 65*, 23–48.

Durham, A. G. and A. R. Gallant (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes (with discussion). *J. Business and Economic Statist. 20*, 297–316.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica 50*, 987–1007.

Engle, R. F. and J. R. Russell (1998). Forecasting transaction rates: the autoregressive conditional duration model. *Econometrica 66*, 1127–62.

Engle, R. F. and M. W. Watson (1981). A one-factor multivariate time series model of metropolitan wage rates. *J. American Statistical Association 76*, 774–81.

Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalised linear models. *J. American Statistical Association 87*, 501–9.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Berlin: Springer.

Fessler, J. A. (1991). Nonparametric fixed-interval smoothing with vector splines. *IEEE Trans. Signal Process. 39*, 852–59.

Fletcher, R. (1987). *Practical Methods of Optimization* (2nd ed.). New York: John Wiley.

Francke, M. K., S. J. Koopman, and A. F. de Vos (2010). Likelihood functions for state space models with diffuse initial conditions. *Journal of Time Series Analysis 31*, 407–14.

Fraser, D. and J. Potter (1969). The optimum linear smoother as a combination of two optimum linear filters. *IEEE Transactions on Automatic Control 4*, 387–90.

French, K. R., G. W. Schwert, and R. F. Stambaugh (1987). Expected stock returns and volatility. *J. Financial Economics 19*, 3–29. Reprinted as pp. 61–86 in Engle, R. F. (1995), *ARCH: Selected Readings*, Oxford: Oxford University Press.

Fridman, M. and L. Harris (1998). A maximum likelihood approach for non-gaussian stochastic volatility models. *J. Business and Economic Statist. 16*, 284–91.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Analysis 15*, 183–202.

Frühwirth-Schnatter, S. (2004). Efficient Bayesian parameter estimation. In A. C. Harvey, S. J. Koopman, and N. Shephard (Eds.), *State Space and Unobserved Components Models*. Cambridge: Cambridge University Press.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York, NY: Springer Press.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika 85*, 215–27.

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulations for Bayesian Inference* (2nd ed.). London: Chapman and Hall.

Gelfand, A. E. and A. F. M. Smith (Eds.) (1999). *Bayesian Computation.* Chichester: John Wiley and Sons.

Gelman, A. (1995). Inference and monitoring convergence. In W. K. Gilks, S. Richardson and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 131–43. London: Chapman & Hall.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis.* London: Chapman & Hall.

Geweke, J. (1977). The dynamic factor analysis of economic time series. In D. J. Aigner and A. S. Goldberger (Eds.), *Latent Variables in Socio-economic Models.* Amsterdam: North-Holland.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–39.

Ghysels, E., A. C. Harvey, and E. Renault (1996). Stochastic volatility. In C. R. Rao and G. S. Maddala (Eds.), *Statistical Methods in Finance*, pp. 119–91. Amsterdam: North-Holland.

Gilks, W. K., S. Richardson, and D. J. Spiegelhalter (Eds.) (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics 31*, 1208–12.

Golub, G. H. and C. F. Van Loan (1996). *Matrix Computations* (3rd ed.). Baltimore: The Johns Hopkins University Press.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). A novel approach to non-linear and non-Gaussian Bayesian state estimation. *IEE-Proceedings F 140*, 107–13.

Granger, C. W. J. and R. Joyeau (1980). An introduction to long memory time series models and fractional differencing. *J. Time Series Analysis 1*, 15–39.

Granger, C. W. J. and P. Newbold (1986). *Forecasting Economic Time Series* (2nd ed.). Orlando: Academic Press.

Green, P. and B. W. Silverman (1994). *Nonparameteric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman & Hall.

Hamilton, J. (1994). *Time Series Analysis.* Princeton: Princeton University Press.

Hammersley, J. M. and D. C. Handscomb (1964). *Monte Carlo Methods.* London: Methuen and Co.

Hammersley, J. M. and K. W. Morton (1954). Poor man's Monte Carlo. *J. Royal Statistical Society B 16*, 23–38.

Handschin, J. (1970). Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica 6*, 555–63.

Handschin, J. and D. Q. Mayne (1969). Monte Carlo techniques to estimate the conditional expectations in multi-stage non-linear filtering. *International Journal of Control 9*, 547–59.

Hardle, W. (1990). *Applied Nonparameteric Regression.* Cambridge: Cambridge University Press.

Harrison, J. and C. F. Stevens (1976). Bayesian forecasting (with discussion). *J. Royal Statistical Society B 38*, 205–47.

Harrison, J. and M. West (1991). Dynamic linear model diagnostics. *Biometrika 78*, 797–808.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A. C. (1993). *Time Series Models* (2nd ed.). Hemel Hempstead: Harvester Wheatsheaf.

Harvey, A. C. (1996). Intervention analysis with control groups. *International Statistical Review 64*, 313–28.

Harvey, A. C. (2006). Forecasting with unobserved components time series models. In G. Elliot, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, pp. 327–412. Amsterdam: Elsevier Science Publishers.

Harvey, A. C. and C.-H. Chung (2000). Estimating the underlying change in unemployment in the UK (with discussion). *J. Royal Statistical Society A 163*, 303–39.

Harvey, A. C. and J. Durbin (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling, (with discussion). *J. Royal Statistical Society A 149*, 187–227.

Harvey, A. C. and C. Fernandes (1989). Time series models for count data or qualitative observations. *J. Business and Economic Statist. 7*, 407–17.

Harvey, A. C. and S. J. Koopman (1992). Diagnostic checking of unobserved components time series models. *J. Business and Economic Statist. 10*, 377–89.

Harvey, A. C. and S. J. Koopman (1997). Multivariate structural time series models. In C. Heij, H. Schumacher, B. Hanzon, and C. Praagman (Eds.), *Systematic Dynamics in Economic and Financial Models*, pp. 269–98. Chichester: John Wiley and Sons.

Harvey, A. C. and S. J. Koopman (2000). Signal extraction and the formulation of unobserved components models. *Econometrics Journal 3*, 84–107.

Harvey, A. C. and S. J. Koopman (2009). Unobserved components models in economics and finance. *IEEE Control Systems Magazine 29*, 71–81.

Harvey, A. C. and S. Peters (1984). Estimation procedures for structural time series models. Discussion paper, London School of Economics.

Harvey, A. C. and G. D. A. Phillips (1979). The estimation of regression models with autoregressive-moving average disturbances. *Biometrika 66*, 49–58.

Harvey, A. C., E. Ruiz, and N. Shephard (1994). Multivariate stochastic variance models. *Rev. Economic Studies 61*, 247–64.

Harvey, A. C. and N. Shephard (1990). On the probability of estimating a deterministic component in the local level model. *J. Time Series Analysis 11*, 339–47.

Harvey, A. C. and N. Shephard (1993). Structural time series models. In G. S. Maddala, C. R. Rao, and H. D. Vinod (Eds.), *Handbook of Statistics, Volume 11*. Amsterdam: Elsevier Science Publishers.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. London: Chapman & Hall.

Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. Research memorandum, Carnegie Institute of Technology, Pittsburgh, Pennsylvania.

Hull, J. and A. White (1987). The pricing of options on assets with stochastic volatilities. *J. Finance 42*, 281–300.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press [Reprinted by Dover in 2007].

Johansen, A. M. and A. Doucet (2008). A note on auxiliary particle filters. *Statistics and Probability Letters 78*(12), 1498–1504.

Jones, R. H. (1993). *Longitudinal Data with Serial Correlation: A State-space approach*. London: Chapman & Hall.

Journel, A. (1974). Geostatistics for conditional simulation of ore bodies. *Economic Geology 69*, 673–687.

Julier, S. J. and J. K. Uhlmann (1997). A new extension of the Kalman filter to nonlinear systems. In I. Kadar (Ed.), *Signal Processing, Sensor Fusion, and Target Recognition VI*, Volume 3068, pp. 182–193.

Jungbacker, B. and S. J. Koopman (2005). Model-based measurement of actual volatility in high-frequency data. In T. B. Fomby and D. Terrell (Eds.), *Advances in Econometrics*. New York: JAI Press.

Jungbacker, B. and S. J. Koopman (2006). Monte Carlo likelihood estimation for three multivariate stochastic volatility models. *Econometric Reviews 25*, 385–408.

Jungbacker, B. and S. J. Koopman (2007). Monte Carlo estimation for nonlinear non-Gaussian state space models. *Biometrika 94*, 827–39.

Jungbacker, B. and S. J. Koopman (2008). Likelihood-based analysis for dynamic factor models. Discussion paper Vrije Universiteit, Amsterdam.

Kahn, H. and A. W. Marshall (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of the Operational Research Society of America 1*, 263–271.

Kailath, T. and P. Frost (1968). An innovations approach to least-squares estimation. part ii: linear smoothing in additive white noise. *IEEE Transactions on Automatic Control 13*, 655–60.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Engineering, Transactions ASMA, Series D 82*, 35–45.

Kim, C. J. and C. R. Nelson (1999). *State Space Models with Regime Switching.* Cambridge, Massachusetts: MIT Press.

Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics 46*, 605–23.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Computational and Graphical Statistics 5*, 1–25.

Kitagawa, G. and W. Gersch (1996). *Smoothness Priors Analysis of Time Series.* New York: Springer Verlag.

Kloek, T. and H. K. Van Dijk (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica 46*, 1–20.

Kohn, R. and C. F. Ansley (1989). A fast algorithm for signal extraction, influence and cross-validation. *Biometrika 76*, 65–79.

Kohn, R., C. F. Ansley, and C.-M. Wong (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika 79*, 335–46.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika 80*, 117–26.

Koopman, S. J. (1997). Exact initial Kalman filtering and smoothing for nonstationary time series models. *J. American Statistical Association 92*, 1630–38.

Koopman, S. J. (1998). Kalman filtering and smoothing. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*. Chichester: Wiley and Sons.

Koopman, S. J. and C. S. Bos (2004). State space models with a common stochastic variance. *J. Business and Economic Statist. 22*, 346–57.

Koopman, S. J. and J. Durbin (2000). Fast filtering and smoothing for multivariate state space models. *J. Time Series Analysis 21*, 281–96.

Koopman, S. J. and J. Durbin (2003). Filtering and smoothing of state vector for diffuse state space models. *J. Time Series Analysis 24*, 85–98.

Koopman, S. J. and A. C. Harvey (2003). Computing observation weights for signal extraction and filtering. *J. Economic Dynamics and Control 27*, 1317–33.

Koopman, S. J., A. C. Harvey, J. A. Doornik, and N. Shephard (2010). *Stamp 8.3: Structural Time Series Analyser, Modeller and Predictor*. London: Timberlake Consultants.

Koopman, S. J. and E. Hol-Uspensky (2002). The Stochastic Volatility in Mean model: Empirical evidence from international stock markets. *J. Applied Econometrics 17*, 667–89.

Koopman, S. J., B. Jungbacker, and E. Hol (2005). Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. *J. Empirical Finance 12*, 445–75.

Koopman, S. J. and K. M. Lee (2009). Seasonality with trend and cycle interactions in unobserved components models. *J. Royal Statistical Society C, Applied Statistics 58*, 427–48.

Koopman, S. J. and A. Lucas (2008). A non-Gaussian panel time series model for estimating and decomposing default risk. *J. Business and Economic Statist. 26*, 510–25.

Koopman, S. J., A. Lucas, and A. Monteiro (2008). The multi-state latent factor intensity model for credit rating transitions. *J. Econometrics 142*, 399–424.

Koopman, S. J., A. Lucas, and M. Scharth (2011). Numerically accelerated importance sampling for nonlinear non-Gaussian state space models. Discussion paper, Vrije Universiteit, Amsterdam.

Koopman, S. J., A. Lucas, and B. Schwaab (2011). Modeling frailty-correlated defaults using many macroeconomic covariates. *J. Econometrics 162*, 312–25.

Koopman, S. J., M. Mallee, and M. van der Wel (2010). Analyzing the term structure of interest rates using the dynamic Nelson-Siegel model with time-varying parameters. *J. Business and Economic Statist. 28*, 329–43.

Koopman, S. J., M. Ooms and I. Hindrayanto (2009). Periodic Unobserved Cycles in Seasonal Time Series with an Application to U.S. Unemployment. *Oxford Bulletin of Economics and Statistics 71*, 683–713.

Koopman, S. J. and N. Shephard (1992). Exact score for time series models in state space form. *Biometrika 79*, 823–6.

Koopman, S. J., N. Shephard, and D. D. Creal (2009). Testing the assumptions behind importance sampling. *J. Econometrics 149*, 2–11.

Koopman, S. J., N. Shephard, and J. A. Doornik (1999). Statistical algorithms for models in state space form using SsfPack 2.2. *Econometrics Journal 2*, 113–66. http://www.ssfpack.com/

Koopman, S. J., N. Shephard, and J. A. Doornik (2008). *Statistical Algorithms for Models in State Space Form: SsfPack 3.0.* London: Timberlake Consultants Press.

Lawley, D. N. and A. E. Maxwell (1971). *Factor Analysis as a Statistical Method* (2 ed.). London: Butterworths.

Lee, K. M. and S. J. Koopman (2004). Estimating stochastic volatility models: a comparison of two importance samplers. *Studies in Nonlinear Dynamics and Econometrics 8*, Art 5.

Lehmann, E. (1983). *Theory of Point Estimation.* New York: Springer.

Liesenfeld, R. and R. Jung (2000). Stochastic volatility models: conditional normality versus heavy-tailed distributions. *J. Applied Econometrics 15*, 137–160.

Liesenfeld, R. and J. F. Richard (2003). Univariate and multivariate stochastic volatility models: Estimation and diagnostics. *J. Empirical Finance 10*, 505–531.

Litterman, R. and J. Scheinkman (1991). Common factors affecting bond returns. *Journal of Fixed Income 1*(1), 54–61.

Liu, J. and R. Chen (1998). Sequential Monte Carlo methods for dynamic systems. *J. American Statistical Association 93*, 1032–44.

Ljung, G. M. and G. E. P. Box (1978). On a measure of lack of fit in time series models. *Biometrika 66*, 67–72.

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* New York: Wiley.

Makridakis, S., S. C. Wheelwright, and R. J. Hyndman (1998). *Forecasting: Methods and Applications* (3rd ed.). New York: John Wiley and Sons.

Marshall, A. W. (1956). The use of multi-stage sampling schemes in Monte Carlo computations. In M. Meyer (Ed.), *Symposium on Monte Carlo Methods*, pp. 123–140. New York: Wiley.

Maskell, S. (2004). Basics of the particle filter. In A. C. Harvey, S. J. Koopman, and N. Shephard (Eds.), *State Space and Unobserved Components Models.* Cambridge: Cambridge University Press.

Mayne, D. Q. (1966). A solution of the smoothing problem for linear dynamic systems. *Automatica 4*, 73–92.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

Mills, T. C. (1993). *Time Series Techniques for Economists* (2nd ed.). Cambridge: Cambridge University Press.

Monahan, J. F. (1993). Testing the behaviour of importance sampling weights. *Computer Science and Statistics: Proceedings of the 25th Annual Symposium on the Interface*, 112–117.

Monahan, J. F. (2001). *Numerical Methods of Statistics.* Cambridge: Cambridge University Press.

Morf, J. F. and T. Kailath (1975). Square root algorithms for least squares estimation. *IEEE Transactions on Automatic Control 20*, 487–97.

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *J. American Statistical Association 55*, 299–305.

Nelson, C. R. and A. Siegel (1987). Parsimonious modelling of yield curves. *Journal of Business 60-4*, 473–89.

Nocedal, J. and S. J. Wright (1999). *Numerical Optimization.* New York: Springer Verlag.

Pfeffermann, D. and R. Tiller (2005). Bootstrap Approximation to Prediction MSE for State-Space Models with Estimated Parameters. *J. Time Series Analysis 26*, 893–916.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: auxiliary particle filter. *J. American Statistical Association 94*, 590–99.

Plackett, R. L. (1950). Some theorems in least squares. *Biometrika 37*, 149–57.

Poirier, D. J. (1995). *Intermediate Statistics and Econometrics.* Cambridge: MIT.

Proietti, T. (2000). Comparing seasonal components for structural time series models. *International Journal of Forecasting 16*, 247–60.

Quah, D. and T. J. Sargent (1993). A dynamic index model for large cross sections. In J. H. Stock and M. Watson (Eds.), *Business Cycles, Indicators and Forecasting*, pp. 285–306. Chicago: University of Chicago Press.

Quenneville, B. and A. C. Singh (1997). Bayesian prediction mean squared error for state space models with estimated parameters. *J. Time Series Analysis 21*, 219–36.

Quintana, J. M. and M. West (1987). An analysis of international exchange rates using multivariate DLM's. *The Statistican 36*, 275–81.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.). New York: John Wiley & Sons.

Rauch, H., F. Tung, and C. Striebel (1965). Maximum likelihood estimation of linear dynamic systems. *AIAA Journal 3*, 1445–50.

Ray, B. and R. S. Tsay (2000). Long- range dependence in daily stock volatilities. *J. Business and Economic Statist. 18*, 254–62.

Richard, J. F. and W. Zhang (2007). Efficient high-dimensional importance sampling. *J. Econometrics 141*, 1385–1411.

Ripley, B. D. (1987). *Stochastic Simulation.* New York: Wiley.

Robert, C. and G. Casella (2010). *Introducing Monte Carlo Methods with R.* New York: Springer.

Rosenberg, B. (1973). Random coefficients models: the analysis of a cross-section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement 2*, 399–428.

Rydberg, T. H. and N. Shephard (2003). Dynamics of trade-by-trade price movements: decomposition and models. *Journal of Financial Econometrics 1*, 2–25.

Sage, A. P. and J. L. Melsa (1971). *Estimation Theory with Applications to Communication and Control.* New York: McGraw Hill.

Särkkä, S. and J. Hartikainen (2010). On Gaussian optimal smoothing of non-linear state space models. *IEEE Transactions on Automatic Control 55*, 1038–1941.

Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Transactions on Information Theory 11*, 61–70.

Shephard, N. (1994a). Local scale model: state space alternative to integrated GARCH processes. *J. Econometrics 60*, 181–202.

Shephard, N. (1994b). Partial non-Gaussian state space. *Biometrika 81*, 115–31.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In D. R. Cox, D. V. Hinkley, and O. E. Barndorff-Nielson (Eds.), *Time Series Models in Econometrics, Finance and Other Fields*, pp. 1–67. London: Chapman & Hall.

Shephard, N. (2005). *Stochastic Volatility: Selected Readings.* Oxford: Oxford University Press.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*, 653–67.

Shumway, R. H. and D. S. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Analysis 3*, 253–64.

Shumway, R. H. and D. S. Stoffer (2000). *Time Series Analysis and Its Applications.* New York: Springer-Verlag.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Royal Statistical Society B 47*, 1–52.

Smith, J. Q. (1979). A generalization of the Bayesian steady forecasting model. *J. Royal Statistical Society B 41*, 375–87.

Smith, J. Q. (1981). The multiparameter steady model. *J. Royal Statistical Society B 43*, 256–60.

Snyder, R. D. and G. R. Saligari (1996). Initialization of the Kalman filter with partially diffuse initial conditions. *J. Time Series Analysis 17*, 409–24.

So, M. K. P. (2003). Posterior mode estimation for nonlinear and non-Gaussian state space models. *Statistica Sinica 13*, 255–74.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *J. American Statistical Association 97*, 1167–79.

Stoffer, D. S. and K. D. Wall (1991). Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter. *J. American Statistical Association 86*, 1024–33.

Stoffer, D. S. and K. D. Wall (2004). Resampling in state space models. In A. C. Harvey, S. J. Koopman, and N. Shephard (Eds.), *State Space and Unobserved Components Models*. Cambridge: Cambridge University Press.

Taylor, S. J. (1986). *Modelling Financial Time Series*. Chichester: John Wiley.

Teräsvirta, T., D. Tjostheim, and C. W. J. Granger (2011). *Modelling Nonlinear Economic Time Series*. Oxford: Oxford University Press.

Theil, H. and S. Wage (1964). Some observations on adaptive forecasting. *Management Science 10*, 198–206.

Tsiakas, I. (2006). Periodic stochastic volatility and fat tails. *J. Financial Econometrics 4*, 90–135.

Valle e Azevedo, J., S. J. Koopman, and A. Rua (2006). Tracking the business cycle of the Euro area: a multivariate model-based bandpass filter. *J. Business and Economic Statist. 24*, 278–90.

van der Merwe, R., A. Doucet, and N. de Freitas (2000). The unscented particle filter. In T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.), *Advances in Neural Information Processing Systems*, pp. 13. Cambridge: MIT Press.

Wahba, G. (1978). Improper priors, spline smoothing, and the problems of guarding against model errors in regression. *J. Royal Statistical Society B 40*, 364–72.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Watson, M. W. and R. F. Engle (1983). Alternative algorithms for the estimation of dynamic factor, MIMIC and varying coefficient regression. *J. Econometrics 23*, 385–400.

Wecker, W. E. and C. F. Ansley (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. American Statistical Association 78*, 81–9.

West, M. and J. Harrison (1989). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models* (2nd ed.). New York: Springer-Verlag.

West, M., J. Harrison, and H. S. Migon (1985). Dynamic generalised models and Bayesian forecasting (with discussion). *J. American Statistical Association 80*, 73–97.

Whittle, P. (1991). Likelihood and cost as path integrals (with discussion). *J. Royal Statistical Society B 53*, 505–38.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science 6*, 324–42.

Yee, T. W. and C. J. Wild (1996). Vector generalized additive models. *J. Royal Statistical Society B 58*, 481–93.

Young, P. C. (1984). *Recursive Estimation and Time Series Analysis*. New York: Springer-Verlag.

Young, P. C., K. Lane, C. N. Ng, and D. Palmer (1991). Recursive forecasting, smoothing and seasonal adjustment of nonstationary environmental data. *J. of Forecasting 10*, 57–89.

Yu, J. (2005). On leverage in a stochastic volatility model. *J. Econometrics 127*, 165–78.

Author Index

Subject Index